UNIVERSIDADE FEDERAL DE ALAGOAS-UFAL
INSTITUTO DE COMPUTAÇÃO / IC
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

WELLIGNTON BATISTA DA SILVA

RECOMENDAÇÃO DE UM MODELO DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE RISCO CARDIOVASCULAR COM BIOMARCADORES DA SÍNDROME METABÓLICA E ESCORE DE FRAMINGHAM

MACEIÓ
2023

Wellignton Batista da Silva

RECOMENDAÇÃO DE UM MODELO DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE RISCO CARDIOVASCULAR COM BIOMARCADORES DA SÍNDROME METABÓLICA E ESCORE DE FRAMINGHAM

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas - UFAL, INSTITUTO DE COMPUTAÇÃO / IC .

Orientador: Prof. Dr. RAFAEL DE AMORIM SILVA

Maceió
2023

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Helena Cristina Pimentel do Vale CRB4 - 661

S586r Silva, Wellignton Batista da.
Recomendação de um modelo de aprendizado de máquina parapredição de risco cardiovascular com biomarcadores da síndrome metabólica e escore de framingham / Wellignton Batista. – 2023.
135 f. : il.

Orientador: Rafael de Amorim Silva.
Dissertação(mestrado em Informática) – Universidade Federal de Alagoas, Instituto de Computação. Maceió, 2023.

Bibliografia: f. 92-102.
Apêndices: f. 103-135.

1. Aprendizado de máquina. 2. Modelo de predição. 3. Síndrome metabólica. 5. Escore de risco global (ERG) de framingham. I. Título.

CDU: 004.78:61

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PPGI
Instituto de Computação/UFAL
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Folha de Aprovação

WELLIGNTON BATISTA DA SILVA

RECOMENDAÇÃO DE UM MODELO DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE RISCO CARDIOVASCULAR COM BIOMARCADORES DA SÍNDROME METABÓLICA E ESCORE DE FRAMINGHAM

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 30 de agosto de 2023.

Banca Examinadora:

Documento assinado digitalmente
gov.br RAFAEL DE AMORIM SILVA
Data: 30/08/2023 16:24:12-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. RAFAEL DE AMORIM SILVA
UFAL – Instituto de Computação
Orientador

Documento assinado digitalmente
gov.br ALMIR PEREIRA GUIMARAES
Data: 31/08/2023 21:43:05-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. ALMIR PEREIRA GUIMARÃES
UFAL – Instituto de Computação
Examinador Externo

Documento assinado digitalmente
gov.br BRUNO ALMEIDA PIMENTEL
Data: 31/08/2023 19:31:15-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. BRUNO ALMEIDA PIMENTEL
UFAL – Instituto de Computação
Coorientador

Documento assinado digitalmente
gov.br RAFAEL FERREIRA LEITE DE MELLO
Data: 31/08/2023 08:34:57-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. RAFAEL FERREIRA LEITE DE MELLO
UFRPE-Universidade Federal Rural de Pernambuco
Examinador Externo

Documento assinado digitalmente
gov.br RANILSON OSCAR ARAUJO PAIVA
Data: 31/08/2023 17:58:58-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. RANILSON OSCAR ARAÚJO PAIVA
UFAL – Instituto de Computação
Examinador Interno

AGRADECIMENTOS

Gostaria de expressar minha imensa gratidão à Nossa Senhora Virgem dos Pobres por sua presença constante e por trazer paz à minha mente durante os momentos difíceis enfrentados ao longo deste trabalho e em outras áreas da minha vida. Também quero expressar minha profunda gratidão ao Nosso Senhor Jesus Cristo, cujos ensinamentos sobre a simplicidade de viver têm sido uma fonte constante de inspiração para mim.

Gostaria de estender meus mais sinceros agradecimentos ao Prof. Dr. Rafael Amorim, meu orientador, que, com sua sabedoria, soube me guiar e acreditar que, com dedicação, seria possível concluir esta tarefa. Também desejo expressar minha gratidão ao coorientador, o Prof. Dr. Bruno Almeida, pelo seu valioso apoio e contribuição. Além disso, reconheço a importância de outras pessoas nesse processo, como o Prof. Dr. Alan Pedro e o Prof. Dr. Fernando Pimentel, que foram meus primeiros orientadores e contribuíram para o início dessa jornada. Sou grato também a todos os professores que conduziram suas aulas de maneira magnífica e aos avaliadores da qualificação, que contribuíram para o desenvolvimento acadêmico e relevante da dissertação.

Por último, mas não menos importante, quero expressar minha profunda gratidão à minha maravilhosa família e aos meus amigos. Peço desculpas pelos momentos em que estive ausente, dedicando-me à pesquisa e aos experimentos. Seu apoio incondicional e compreensão foram fundamentais para o meu progresso e sucesso nessa empreitada.

Mais uma vez, expresso minha gratidão a todos que contribuíram para minha jornada acadêmica e pessoal. Sou verdadeiramente abençoado por ter tido o apoio e a orientação de pessoas tão maravilhosas ao longo desse caminho.

RESUMO

A previsão de eventos cardiovasculares em pacientes diagnosticados com Síndrome Metabólica (SM) é um tema de grande relevância para a área da Saúde, em geral, e fundamental para a da Endocrinologia. Esta dissertação objetiva recomendar um modelo de Machine Learning (ML) para estimar os riscos de eventos cardiovasculares em pacientes com SM, explorando os marcadores do escore de Framingham (FRS) e da SM. Metodologicamente, utilizamos um modelo de regressão logística e análises com árvore de decisão, floresta aleatória, impulsionamento de gradiente, máquina de vetores de suporte e k-vizinhos mais próximos para testar nossa hipótese central de que os bioindicadores (variáveis relacionadas à SM) exercem um impacto positivo, forte e significativo nos eventos cardiovasculares em pacientes com SM. Tecnicamente a pesquisa foi conduzida por meio de experimentos realizados em diferentes cenários. No primeiro cenário, um algoritmo foi desenvolvido para avaliar o risco cardiovascular em pacientes com e sem SM. Nos cenários 2, 3 e 4, foram analisados pacientes com e sem SM, considerando os marcadores de SM e FRS como variáveis dependentes, enquanto a condição de Síndrome Metabólica foi adotada como variável independente. No quinto cenário, uma análise foi realizada para selecionar o modelo de regressão e classificação mais adequado para a predição do risco cardiovascular em um conjunto de dados combinado de doenças cardíacas. No sexto cenário, o modelo desenvolvido foi fundamentado no Escore de Risco Global (ERG) de Framingham, incorporando os marcadores da SM nos experimentos. Os dados foram obtidos a partir do repositório do *National Center for Health Statistics* (NHANES), um conjunto de dados combinados de doenças cardíacas do Repositório de aprendizado de máquina UCI. e da plataforma Kaggle. Em resumo, os principais achados desta dissertação são os seguintes: (1) No primeiro cenário, uma diferença percentual de 81,74% foi observada nas médias de Risco ECV entre as populações com e sem Síndrome Metabólica, evidenciando um aumento significativo do risco cardiovascular na população com SM; (2) nos cenários dois, três e quatro, o modelo Random Forest (RF) se destacou, alcançando alta acurácia em todas as combinações de variáveis, especialmente na combinação dos marcadores da SM com o marcador de sexo; (3) no quinto cenário, o modelo RF foi identificado como o mais indicado, destacando a importância das variáveis relacionadas à SM na predição do risco cardiovascular e ressaltando a necessidade de aprimoramentos nos modelos para melhor identificação dos casos positivos; (4) tanto o modelo com quatro marcadores da SM quanto o modelo com cinco marcadores da SM combinados ao escore de Framingham (SM + FRS) apresentaram desempenho considerável, com correlações e acurácias significativas (0.80 e 0.84, respectivamente). Essas combinações mais simples de variáveis podem ser uma abordagem interessante, uma vez que fornecem informações relevantes para a predição do risco cardiovascular de forma menos invasiva, evitando a necessidade de exames mais complexos.

Palavras-chave: Aprendizado de Máquina; Modelo de Predição; Síndrome Metabólica; Risco Vascular; Escore de Risco Global (ERG) de Framingham.

ABSTRACT

The prediction of cardiovascular events in patients diagnosed with Metabolic Syndrome (MS) is a topic of great importance in the field of Health in general and crucial for Endocrinology. This dissertation aims to recommend a Machine Learning (ML) model to estimate the risks of cardiovascular events in patients with MS, exploring the markers of the Framingham Risk Score (FRS) and MS. Methodologically, we used a logistic regression model and conducted analyses with decision trees, random forest, gradient boosting, support vector machine, and k-nearest neighbors to test our central hypothesis that bioindicators (variables related to MS) exert a positive, strong, and significant impact on cardiovascular events in MS patients. Technically, the research was carried out through experiments conducted in different scenarios. In the first scenario, an algorithm was developed to assess cardiovascular risk in patients with and without MS. In scenarios 2, 3, and 4, patients with and without MS were analyzed, considering MS and FRS markers as dependent variables, while the Metabolic Syndrome condition was adopted as an independent variable. In the fifth scenario, an analysis was performed to select the most suitable regression and classification model for predicting cardiovascular risk in a combined dataset of heart diseases. In the sixth scenario, the developed model was based on the Framingham Global Risk Score (FRS), incorporating MS markers into the experiments. Data were obtained from the National Center for Health Statistics (NHANES) repository, a combined dataset of heart diseases from the UCI Machine Learning Repository, and the Kaggle platform. In summary, the main findings of this dissertation are as follows: In the first scenario, a percentage difference of 81.74% was observed in mean CVD Risk between populations with and without Metabolic Syndrome, demonstrating a significant increase in cardiovascular risk in the population with MS. In scenarios two, three, and four, the Random Forest (RF) model excelled, achieving high accuracy in all variable combinations, especially in the combination of MS markers with gender. In the fifth scenario, the RF model was identified as the most suitable, emphasizing the importance of variables related to MS in predicting cardiovascular risk and highlighting the need for model improvements to better identify positive cases. Both the model with four MS markers and the model with five MS markers combined with the Framingham Risk Score (SM + FRS) demonstrated considerable performance, with significant correlations and accuracies (0.80 and 0.84, respectively). These simpler variable combinations can be an interesting approach as they provide relevant information for predicting cardiovascular risk in a less invasive manner, avoiding the need for more complex tests.

Keywords: Machine Learning; Cardiovascular Event Prediction; Metabolic Syndrome; Framingham Global Risk Score (GRS).

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AHA - American Heart Association

ADASYN - Adaptative Synthetic Sampling Approach for Imbalanced Datasets (Enfoque de muestreo sintético adaptativo

ARPA - Agência de Pesquisa de Projetos Avançados

AUC - área sob a curva ROC

AVC: - Acidente Vascular Cerebral

ALT - Alanina Aminotransferase

AST - Aspartato Aminotransferase

AUC - Área sob a Curva

CPK - Creatina Quinase

CRM - Conselho Regional de Medicina

CHD - Doença Coronariana Futura

DAC - Doença Arterial Coronariana

DCV - Doenças Cardiovasculares

DT - Árvore de Decisão -)Decision Tree)

DTC - Decision Tree Classification

DTR - Decision Tree Regressor

DM - Diabetes Mellitus

EHR - Registro Eletrônico de Saúde (Electronic Health Record)

ECG - Eletrocardiograma

FPR - Falsos Positivos

FRS - Framingham Heart Failure Risk Score

FSCNCA - Feature selection using neighborhood component analysis (Selección de características mediante el análisis de componentes de vecindad)

GGT - Gama-Glutamiltransferase

HDL - Lipoproteína de alta densidade (do inglês High-Density Lipoprotein)

HOMA-IR - Avaliação do Modelo Homeostático de Resistência à Insulina (Homeostatic Model Assessment of Insulin Resistance)

IA - Inteligência Artificial

IMC - Índice de Massa Corporal

KNN - (K-Nearest Neighbors): K-Vizinhos-Mais-Próximos

LR - Regressão Linear (Linear Regression)

Logistic Regression - Regressão Logística

MAE - Mean Absolute Error

ML - Machine Learning (Aprendizado de Máquina)

MLAs - Algoritmos de Aprendizado de Máquina em Conjunto

MSE - Mean Squared Error

Marcadores SM - Marcadores da Síndrome Metabólica

Marcadores SM + Sexo - Marcadores da Síndrome Metabólica combinados com o Sexo

Marcadores SM + Idade - Marcadores da Síndrome Metabólica combinados com a Idade

Marcadores SM + Obesidade - Marcadores da Síndrome Metabólica combinados com a Obesidade

Marcadores SM + FRS - Marcadores da Síndrome Metabólica combinados com o escore de Framingham

Marcadores FRS - Marcadores do escore de Framingham

Marcadores FRS + Obesidade - Marcadores do escore de Framingham combinados com a Obesidade

Marcadores FRS + SM + Obesidade - Marcadores do escore de Framingham combinados com os marcadores da Síndrome Metabólica e a Obesidade

MS - Metabolic Syndrome - (SM - Síndrome Metabólica)

oldpeak - Depressão do segmento ST induzida pelo exercício

NHANES - Pesquisa Nacional de Avaliação de Saúde e Nutrição (National Health and Nutritional Evaluation Survey)

NHLBI - National Heart, Lung, and Blood Institute

NCEP-ATP III - National Cholesterol Education Program's Adult Treatment Panel III

OMS - Organização Mundial da Saúde

RQE - Registro de Qualificação de Especialista

RF - Floresta Aleatória (Random Forest)

RC - Risco Cardiovascular

RSL - Revisão Sistemática da Literatura

RMSE - Root Mean Squared Error

SVR - Regressão de Vetor de Suporte (Support Vector Regression)

TPR - Verdadeiros Positivos

SUMÁRIO

1 INTRODUÇÃO

A SM é uma condição clínica que aumenta o risco de desenvolver DCV, sendo a DCV uma das principais causas de mortalidade em todo o mundo. Segundo o *National Heart, Lung, and Blood Institute 2021*, a SM é caracterizada pela presença de três ou mais fatores de risco, como hipertensão arterial, níveis elevados de triglicerídeos, baixos níveis de colesterol HDL, obesidade abdominal e resistência à insulina. A identificação precoce de pacientes em risco é crucial para uma intervenção mais rápida e eficaz. Conforme mencionado por (KALIL et al., 2017), a SM não é uma doença em si, mas um termo que descreve a presença simultânea de vários fatores de RDC em uma pessoa.

Com base nisso, a pesquisa tem como propósito contribuir para o avanço na área da prevenção e gerenciamento de SM e DCV, permitindo a identificação precoce de indivíduos em risco, adoção de medidas preventivas e intervenções clínicas adequadas. A avaliação preditiva pode ser uma ferramenta valiosa na área da saúde, permitindo a construção do perfil dos pacientes a partir do histórico médico, identificação de regiões com maior incidência de patologias específicas, previsão dos custos de exames e internações, a previsão de taxas de ocupação de leitos a aplicação da medicina preventiva para prevenir infecções e doenças potenciais. Essa abordagem visa melhorar o cuidado sobre a qualidade de vida dos pacientes, além de auxiliar no planejamento estratégico das instituições de saúde.

Diante dessa necessidade de previsão e gerenciamento de riscos na área da saúde, surge a pergunta problema desta pesquisa: Como os modelos de ML podem ser aplicados para prever e diagnosticar o RDC em relação à SM, considerando fatores clínicos e comportamentais, a fim de auxiliar a prática médica e outros profissionais envolvidos na prevenção e gerenciamento da SM e das DCVs?

Essa pergunta problema direciona o objetivo central desta pesquisa, que é explorar o potencial dos modelos de ML na previsão e diagnóstico do RDC em relação à SM. Buscar-se-á construir um modelo robusto que incorpore biomarcadores e critérios diagnósticos relevantes, considerando tanto os fatores clínicos quanto os comportamentais. Ao responder a essa pergunta, espera-se fornecer informações valiosas para os profissionais de saúde, contribuindo para a prevenção e o cuidado individualizado dos pacientes, além de destacar a importância do uso de ML na prática clínica.

Para alcançar o objetivo proposto, serão adotados os seguintes passos: em primeiro lugar, serão utilizados os biomarcadores definidos pelos Critérios Diagnósticos para a SM segundo o

NCEP-ATP III, que permitem diagnosticar se um paciente apresenta ou não a SM. Esses critérios são descritos na I Diretriz Brasileira de Diagnóstico e Tratamento da SM (BRANDÃO, 2005), que fornece diretrizes atualizadas para o diagnóstico e tratamento da SM. Através da análise desses biomarcadores, será possível identificar os principais fatores de risco e compreender a relação entre a SM e as DCV. Em seguida, serão considerados os biomarcadores estudados no *Framingham Heart Failure Risk Score* (FRS), que é um escore de risco desenvolvido pelo *Framingham Heart Study* para avaliar o risco de insuficiência cardíaca (National Heart, Lung, and Blood Institute, 1948). Esse escore incorpora diversos fatores de risco, como idade, sexo, pressão arterial, colesterol, entre outros, e permite uma avaliação mais precisa e detalhada do risco de SM e DCV em pacientes.

A utilização de biomarcadores e critérios diagnósticos relevantes no desenvolvimento do modelo de predição proporcionará uma abordagem abrangente e fundamentada para prever e diagnosticar o RDC associado à SM. Essas referências científicas e diretrizes reconhecidas no campo da cardiologia e da saúde cardiovascular respaldam a relevância do uso de biomarcadores e modelos de predição para auxiliar a prática médica como também outros profissionais envolvidos no gerenciamento da SM e das DCV. Ao considerar essas referências e adotar uma abordagem baseada em evidências, busca-se fornecer informações valiosas para os profissionais de saúde, contribuindo para a prevenção, diagnóstico precoce e tratamento adequado dessas condições.

O objetivo geral deste trabalho é explorar a precisão dos modelos de ML na predição de DCV, estabelecendo uma relação entre os marcadores da SM e o FRS. Através da aplicação de técnicas avançadas de ML e com base em uma sólida base de dados, o modelo desenvolvido terá o potencial de fornecer informações valiosas para os profissionais de saúde na tomada de decisões clínicas.

Além do objetivo geral, são estabelecidos objetivos específicos para direcionar a pesquisa. Esses objetivos incluem: identificação e seleção dos biomarcadores relevantes para a predição de DCV relacionadas à SM, análise e processamento dos dados coletados a partir desses marcadores, implementação de algoritmos de ML para o desenvolvimento do modelo preditivo e avaliação da acurácia e eficácia desse modelo. Através da consecução desses objetivos específicos, espera-se contribuir para o avanço na área da prevenção e gerenciamento de DCV, permitindo a identificação precoce de indivíduos em risco e a adoção de medidas preventivas e de intervenções clínicas adequadas. O uso de técnicas de ML, aliado a uma abordagem baseada em evidências e dados confiáveis, tem o potencial de otimizar a prática médica, melhorar a tomada de decisões e, consequentemente, impactar positivamente a saúde e o bem-estar dos pacientes.

A importância deste trabalho se reflete em diferentes âmbitos. Em primeiro lugar, para a sociedade como um todo, a pesquisa busca contribuir para a prevenção e gerenciamento de DCV, que representam uma das principais causas de morbidade e mortalidade em todo o mundo. Ao desenvolver um modelo preditivo preciso e confiável, baseado em técnicas de ML, é possível identificar indivíduos em risco e adotar medidas preventivas mais eficazes, resultando em melhores resultados de saúde e qualidade de vida. No contexto da comunidade científica, este trabalho busca preencher uma lacuna de conhecimento ao explorar a aplicação de modelos de ML na predição de DCV relacionadas à SM (BRANDÃO, 2005). Com a crescente disponibilidade de dados médicos e avanços tecnológicos, é fundamental investigar e avaliar a eficácia dessas técnicas para melhorar a prática clínica e impulsionar a pesquisa na área da saúde.

Ao fornecer informações valiosas e predições precisas sobre o risco de desenvolver DCV, os profissionais de saúde poderão tomar decisões mais embasadas, intervenções, oferecer um cuidado mais direcionado e eficiente aos pacientes. Por sua vez, os pacientes podem se beneficiar com uma detecção precoce de riscos, prevenção adequada e adoção de medidas para um estilo de vida saudável.

Para a coleta de dados, serão empregadas diversas bases de dados relevantes. A primeira delas consiste em informações de pacientes diagnosticados com SM, obtidas a partir do repositório do *National Center for Health Statistics* (NHANES). Essa base de dados é uma fonte valiosa de informações relacionadas a essa condição clínica, contribuindo significativamente para o estudo em questão. Além disso, será utilizado um conjunto de dados combinados de DCV, que reúne informações de pacientes provenientes de diferentes localidades, incluindo Cleveland, Hungria, Suíça e VA Long Beach. Essa abordagem permite uma ampla abrangência e diversidade na análise dos fatores de risco e padrões relacionados às DCV (JANOSI, 1988).

Outro conjunto de dados utilizado é o denominado FRS, disponível no Kaggle, que faz referência ao Estudo de Framingham. Esse estudo é reconhecido como um estudo epidemiológico de longo prazo realizado em Framingham, Massachusetts, nos Estados Unidos, com o propósito de investigar os fatores de risco associados às DCV.

Após a conclusão da etapa de coleta de dados, foram conduzidos diversos experimentos com o objetivo de explorar os marcadores da SM e do FRS. Inicialmente, procedeu-se à análise das 29 variáveis presentes no conjunto de dados extraído do (NHANES), com especial atenção para a identificação dos marcadores relacionados ao FRS e à SM. A partir dessas informações, foi desenvolvido um algoritmo capaz de atribuir pontuações para o RDC de acordo com o modelo de FRS, gerando assim uma nova coluna no conjunto de dados contendo os valores correspondentes

ao RDC estimado para os próximos 10 anos.

Após essa etapa, foram conduzidos testes com o objetivo de classificar os riscos em categorias de baixo, moderado, alto, utilizando os percentuais calculados, verificou-se que a população diagnosticada com SM apresenta o dobro da probabilidade de desenvolver RDC em comparação com a população sem essa condição. Esses testes são de extrema importância, pois fornecerão informações valiosas sobre o RDC e a relação entre os marcadores da SM e do FRS.

Além disso, foram explorados diferentes cenários para aprimorar os modelos utilizados. Esses cenários incluem avaliação do desempenho dos modelos com os marcadores da SM, a utilização dos resultados dos modelos com os marcadores de FRS e combinação dos marcadores da SM e do FRS. A análise desses cenários proporcionou uma compreensão mais abrangente das relações entre os marcadores e permitiu a identificação de abordagens mais efetivas para a prevenção e gerenciamento da SM e das DCV.

Após a coleta dos dados e realização dos experimentos, foram aplicadas técnicas de ML para análise e tratamento dos dados, utilizando as bibliotecas de ML disponíveis em Python. Isso inclui a aplicação de algoritmos de ML para realizar análises quantitativas, como regressão, classificação e agrupamento. Foram utilizadas bibliotecas como scikit-learn, TensorFlow e Keras para implementar e treinar os modelos de ML. Além disso, realizamos análises exploratórias dos dados, utilizando técnicas estatísticas descritivas, como cálculo de média, desvio padrão e análise de correlação. A análise qualitativa será conduzida por meio de técnicas como análise de conteúdo e análise temática, buscando identificar padrões e categorias emergentes nos dados.

O trabalho está organizado em oito capítulos principais, abordando diferentes aspectos da pesquisa. O Capítulo 1 oferece uma introdução ao tema, enquanto o Capítulo 2 justifica a importância do estudo. O Capítulo 3 aborda a fundamentação teórica, incluindo a SM, o RDC e a Inteligência Artificial. O Capítulo 4 apresenta o relato do problema e trabalhos relacionados. A proposta para solucionar o problema é detalhada no Capítulo 5, juntamente com sua fundamentação teórica, hipóteses e relevância. O Capítulo 6 descreve a metodologia utilizada, incluindo o desenvolvimento e escolha do modelo de predição. Os resultados da pesquisa são apresentados no Capítulo 7, com seções dedicadas a diferentes tipos de documentos. O Capítulo 8 consiste na conclusão, resumindo os principais achados e contribuições. Além disso, o trabalho inclui uma seção de referências e apêndices contendo informações adicionais relevantes.

2 JUSTIFICATIVA

A SM é uma condição caracterizada por um conjunto de fatores de risco metabólicos que estão associados a DCVs e outras condições de saúde. A orientação profissional em relação aos pacientes com alto risco de desenvolver um estágio avançado da SM é fundamental para prevenir complicações e melhorar a qualidade de vida. Nesse contexto, se justifica por ser capaz de predizer os riscos para o desenvolvimento de SM, a partir de dado empíricos, permitindo que profissionais de saúde elaborem planos terapêuticos personalizados, visando minimizar a progressão da SM a longo prazo. A hipótese central desta pesquisa é que o modelo proposto será útil não apenas para a prática médica, mas também para a equipe multiprofissional e os pacientes, considerando que a SM tem como principal causa o estilo de vida do indivíduo.

Com base nos dados obtidos, será possível identificar grupos de risco e potenciais ameaças, contribuindo para a construção de perfis de pacientes e o mapeamento de regiões com maior incidência de SM. Além disso, será possível antecipar custos de exames e internações, prever taxas de ocupação de leitos e aplicar medidas de medicina preventiva para evitar infecções e doenças relacionadas.

É importante destacar que a presença da SM e seus componentes também têm sido observados em pacientes com Diabetes Mellitus tipo 1, aumentando o risco de DCV nessa população. Portanto, este trabalho apresentará um modelo de predição de riscos cardiovasculares com base nos biomarcadores utilizados para o diagnóstico de SM e FRS, utilizando técnicas avançadas de ML. Acredita-se que essa abordagem contribuirá para uma prática clínica mais eficiente, permitindo intervenções precoces e personalizadas, resultando em melhores resultados de saúde para os pacientes.

3 FUNDAMENTAÇÃO TEÓRICA

A SM está se tornando cada vez mais comum não só em países desenvolvidos, mas também no Brasil. Embora não haja estudos de prevalência específicos para a população brasileira, estudos em diferentes populações do mundo, como mexicanos, asiáticos e americanos, têm mostrado dados importantes sobre a alta prevalência da SM. As taxas variam de 12,4% a 28,5% em homens e de 10,7% a 40,5% em mulheres, dependendo do critério utilizado e das características da população analisada (JUNQUEIRA et al., 2011).

Devido à associação da SM com um maior número de eventos cardiovasculares, é crucial tratar seus componentes. Um estilo de vida saudável, que inclui não fumar, praticar atividades físicas e perder peso, é fundamental. Em alguns casos, o uso de medicação pode ser necessário, sendo recomendada uma avaliação e orientação específicas de um endocrinologista (METABOLOGIA, 2022). Além disso, outra solução para auxiliar no tratamento é o uso de recursos tecnológicos para predizer o RDC em pacientes diagnosticados com SM.

Devido à sua intrínseca complexidade, a atenção à saúde ainda apresenta incertezas, com mudanças frequentes em protocolos e práticas clínicas. O uso de modelos preditivos baseados em ML tem o potencial de ajudar na tomada de decisão em diferentes momentos da atenção à saúde, especialmente no diagnóstico, intervenção e acompanhamento de problemas de saúde (OBERMEYER; LEE, 2017).

3.1 SM

A SM foi inicialmente descrita pelo diabetologista Gerald Reaven em 1988, durante uma palestra proferida na reunião da *American Diabetes Association* em Nova Orleans. Na ocasião, Reaven apresentou sua teoria sobre a forte ligação entre a resistência à insulina - característica marcante do diabetes tipo 2 - e outras anomalias metabólicas, como a pressão alta e os níveis elevados de triglicerídeos. Sua apresentação, intitulada Banting Lecture, tornou-se lendária entre os especialistas da área. Como parte de sua teoria, Reaven cunhou o termo "Síndrome do Cluster X"para se referir a esse conjunto de condições inter-relacionadas. Vale ressaltar que Reaven é professor de medicina na Universidade de Stanford e sua pesquisa tem sido fundamental para avançarmos na compreensão e no tratamento da SM.

Para (KALIL et al., 2017) a SM não é uma doença em si, mas um termo que designa a presença simultânea de vários fatores de RDC em uma pessoa. Esses fatores de risco são representados por alteração da glicemia (dosagem de glicose no sangue), alteração de triglicérides

(tipo de gordura), alteração de colesterol HDL, pressão alta e acúmulo de gordura na região da barriga (KALIL et al., 2017). Neste sentido, podemos dizer que: a SM é um fator preditivo importante para o desenvolvimento de diabetes no futuro, assim como, predizer o RDC em um paciente diagnosticado com SM no próximos anos.

Por isso, é importante destacar a associação da SM com o RDC em determinado paciente. O aumento da mortalidade geral causada por esta síndrome é cerca de 1,5 vezes na população geral, já na cardiovascular em cerca de 2,5 vezes (NEGRÃO et al., 2005).

De acordo com a (ABESO, 2022), a literatura médica tem mostrado de forma consistente que a presença do diagnóstico de SM, por si só, aumenta a mortalidade geral e cardiovascular em uma determinada população estudada.

A prevenção e o tratamento da SM baseiam-se em medidas como perda de peso e prática regular de atividade física, as quais são consideradas as melhores formas de evitar e tratar essa condição. A detecção precoce do problema pode reduzir o surgimento de doenças cardíacas futuras e proporcionar a oportunidade de promover mudanças no estilo de vida, evitando o desenvolvimento de diversas complicações (SAÚDE, 2017).

Na proposta de pesquisa em questão, que visa realizar experimentos e análises para predizer riscos cardiovasculares com base nos biomarcadores de diagnóstico da SM em pacientes e também os biomarcadores do RFS, é essencial considerar os elementos teóricos apresentados anteriormente. Além disso, é imprescindível que o tema abordado seja igualmente significativo para os interessados, garantindo assim a relevância e o impacto dos resultados obtidos.

Critérios Diagnósticos para a SM segundo o NCEP- ATP III1 e a WHO2 Critério da ATP III Critério da ATP III1.

A presença de 3 ou mais dos seguintes critérios é utilizada para o diagnóstico da SM:

1. Obesidade Abdominal: Cintura > 102 cm em homens e > 88 cm em mulheres
2. Hipertrigliceridemia $\geq$ 150 mg/dl
3. HDL Colesterol Baixo: $\leq$ 40 mg/dl em homens e $\leq$ 50 mg/dl em mulheres
4. Pressão Arterial Elevada: $\geq$ 130/85 mmHg
5. Glicemia de Jejum Elevada: $\geq$ 110 mg/dl

A literatura médica amplamente discute os critérios utilizados para o diagnóstico da SM. De acordo com (ABESO, 2022), o excesso de peso, caracterizado pelo acúmulo de gordura na

circunferência abdominal, é um critério essencial da síndrome. Quando combinado às outras comorbidades relacionadas à resistência à insulina, forma um complexo de fatores de risco que contribuem de forma independente para o desenvolvimento de doença cardiovascular por aterosclerose. É importante salientar que a presença de SM, por si só, aumenta a mortalidade geral e cardiovascular na população estudada.

3.2 RDC

Os primeiros relatos sobre RDC é sobre a aterosclerose, que segundo (LOTUFO, 1999), é uma doença tão antiga quanto a espécie humana. Foi observada em múmias Egípcias datadas do século XV A/C (1) e na China, corpos com mais de 2.000 anos, apresentavam lesões ateroscleróticas das coronárias (LOTUFO, 1999).

Após a descrição da circulação sanguínea por W. Harvey em 1628 (LOTUFO, 1999), W. Heberden, em 1798, descobre a "Angina Pectoris" hoje conhecida como Angina Estável, descrevendo o quadro clínico e reconhecendo a gravidade da doença, porém não tinha ideia da causa desses sintomas (Friesinger, at. al, 1999). No ano seguinte C. Parry (1799) associou a angina à obstrução coronariana, que chamou de "Ossificação".

Em 1849, Vogel fez a primeira referência à presença de colesterol nas placas ateroscleróticas (GIANNINI, 1998). Na época, o patologista mais renomado, R. Virchow (1856), levantou a hipótese de que a aterosclerose era uma doença resultante de perturbações metabólicas na parede das artérias causadas pelo colesterol (II; HURST, 1999). Em 1910, R. Windaus observou que as lesões ateromatosas continham seis vezes mais colesterol livre e vinte vezes mais colesterol esterificado do que a parede normal da artéria (II; HURST, 1999). Apesar das evidências disponíveis até então, ainda não se havia percebido que a hipercolesterolemia era uma das principais causas da aterosclerose.

O estudo pioneiro de A. I. Ignatowski na Academia de Medicina Militar de Leningrado, em 1908, alimentando coelhos com ovos e leite, foi fundamental para a descoberta das placas branco-acinzentadas na aorta dos animais, semelhantes às observadas em corações humanos (CODORNIZ, 2000). No entanto, as principais descobertas sobre os fatores de risco para doenças cardíacas foram realizadas no consagrado estudo de Framingham, uma das primeiras coortes a demonstrar a importância desses fatores (National Heart, Lung, and Blood Institute, 1948). Antes do Framingham, a maioria dos médicos acreditava que a aterosclerose era um processo de envelhecimento inevitável e que a hipertensão arterial era uma consequência fisiológica desse processo, auxiliando o coração a bombear o sangue pelas artérias com lúmen reduzido.

Avançar no entendimento das DCV, incluindo seus aspectos mecanísticos, fisiopatológicos, diagnósticos, prognósticos e terapêuticos, com base sólida na medicina baseada em evidências, é uma tarefa crucial nesse cenário. De acordo com (POLANCZYK, 2005), após cinco décadas de pesquisa sobre os fatores de risco associados às doenças cardíacas, tem sido possível observar em todo o mundo um melhor controle desses fatores e uma redução nas taxas de mortalidade. De fato, dados do estudo de Framingham demonstraram uma redução de 59% na mortalidade por doença coronariana no período de 1950 a 1999 (POLANCZYK, 2005); (National Heart, Lung, and Blood Institute, 1948).

O Ministério da Saúde tem adotado várias iniciativas para reduzir o impacto das doenças não-transmissíveis na população brasileira. Desde o rastreamento de diabete melito em nível nacional, implementação de campanhas sobre hipertensão arterial sistêmica, aplicação de protocolos para manejo agressivo da dislipidemia em coronariopatas, entre outros (POLANCZYK, 2005).

Atualmente, o controle dos níveis de colesterol, pressão arterial e diabetes é fundamental para mitigar doenças cardíacas, vasculares e derrames. É difícil imaginar uma época em que esses e outros fatores de risco não fossem considerados problemas significativos por muitos médicos (National Heart, Lung, and Blood Institute, 1948). É importante reconhecer a evolução do conhecimento médico e a importância de manter uma abordagem baseada em evidências na prevenção e tratamento de DCV.

Destacado na Atualização da Diretriz de Prevenção Cardiovascular da Sociedade Brasileira de Cardiologia de 2019, o Risco Aterosclerótico para SM, Diabetes Mellitus e a Correlação Contínua da Doença Arterial Coronária são de extrema importância para a compreensão das DCV (PRECOMA, 2019). Para (PRECOMA, 2019), a SM e o Diabetes Mellitus representam um espectro de doenças multissistêmicas que afetam o endotélio vascular e contribuem significativamente para a progressão fisiopatológica da DAC.

O RDC (CV) pode aumentar até 20 anos antes do diagnóstico clínico de diabetes mellitus (DM), segundo os critérios atuais. Além disso, a SM é um dos principais fatores de risco para o desenvolvimento de DM e está relacionada a alterações metabólicas que levam à aterotrombose coronariana (CARDIOLOGIA, 2019). Uma metanálise com quase 1 milhão de indivíduos de 87 estudos indicou que a SM está associada a um aumento de duas vezes nos desfechos cardiovasculares e a um aumento de 1,5 vezes na mortalidade por todas as causas, ultrapassando o risco isolado de seus componentes. (CARDIOLOGIA, 2019).

(SOMITI, 2020) afirma que a SM e a sua relação com DCV é caracterizado por um

conjunto de fatores de risco para o coração, relacionados ao acúmulo de gordura abdominal e à resistência à ação da insulina. Entre esses riscos, podemos destacar a dislipidemia, a obesidade centrípeta, a alteração na homeostase glicêmica, isquemia cardíaca e a hipertensão arterial sistêmica. A obesidade central e a gordura visceral, que estão relacionadas à SM, estimulam os processos inflamatórios que se agravam e propiciam doenças vasculares no coração, nas artérias e até no cérebro (SOMITI, 2020).

Nesta pesquisa, serão realizados testes nos algoritmos com o intuito de avaliar o risco de doença cardiovascular. Alguns profissionais da área sugerem que a melhor forma de análise é avaliar cada fator de risco isoladamente, e essa abordagem será adotada neste estudo.

Os resultados indicam que a presença da SM é significante na distribuição do RDC. De fato, a população estudada apresenta um RDC elevado, como evidenciado pela prevalência relevante dos fatores de risco utilizados no FRS e na classificação da SM.

3.3 INTELIGÊNCIA ARTIFICIAL

De acordo com (TUNES, 2022), foi em 1956, durante uma conferência realizada no Dartmouth College, em New Hampshire, Estados Unidos, que o cientista da computação John McCarthy usou pela primeira vez a expressão “inteligência artificial”. As possibilidades eram tão animadoras que órgãos privados e governamentais investiram pesado na área, incluindo a Agência de Pesquisa de Projetos Avançados (ARPA), que é uma organização governamental dos Estados Unidos focada em impulsionar a pesquisa e o desenvolvimento de tecnologias avançadas em várias áreas, incluindo defesa, ciência e tecnologia. Ela é conhecida por financiar projetos inovadores de alto risco que têm o potencial de causar impactos significativos em seus campos de atuação. A ARPA é famosa por seus sucessos no passado, como o desenvolvimento da ARPANET, que foi o precursor da Internet moderna. Sua abordagem se concentra em soluções revolucionárias e no avanço de tecnologias de vanguarda para enfrentar desafios complexos. (KLEINA, 2018).

A inteligência artificial (IA) começou como um campo experimental nos anos 50 com pioneiros como Allen Newell e Herbert Simon, que fundaram o primeiro laboratório de inteligência artificial na Universidade Carnegie Mellon, e McCarty que juntamente com Marvin Minsky, que fundaram o MIT AI Lab em 1959. Foram eles alguns dos participantes na famosa conferência de verão de 1956 em Darthmouth College (BITTENCOURT, 2001).

O campo de pesquisa da inteligência artificial foi estabelecido em uma conferência realizada no campus do Dartmouth College durante o verão de 1956 (KAPLAN, 2022). Os

participantes deste evento se tornaram líderes na pesquisa de IA por muitas décadas.

Um marco importante para o desenvolvimento da inteligência artificial foi o trabalho de Alan Turing intitulado “Computing Machinery and Intelligence” (em português, “Computadores e Inteligência”), escrito em 1950. Esse trabalho contribuiu significativamente para o desenvolvimento do computador moderno e lançou as bases para o estudo da inteligência artificial (LIMA, 2017). O artigo de Turing é dividido em três partes: a primeira aborda o jogo da Imitação e o computador digital, a segunda discute objeções filosóficas à inteligência artificial e, por fim, a terceira parte trata de máquinas que aprendem.

O cientista da computação John McCarthy foi um dos pioneiros da inteligência artificial e foi ele quem cunhou o termo “Inteligência Artificial”. McCarthy também desenvolveu a linguagem de programação LISP, que foi uma das primeiras a ser utilizada no campo da inteligência artificial. A LISP é uma linguagem de programação funcional que se tornou popular na década de 1960 e serviu de base para os estudos da IA. Além disso, a LISP foi utilizada na criação de programas que foram capazes de desafiar a inteligência humana em partidas de xadrez (NOYES, 1992).

Para que a aprendizagem seja significativa, é necessário trabalhar com temas que sejam igualmente relevantes para os envolvidos. Um exemplo disso é a aplicação da IA na área da saúde, uma tecnologia com múltiplas aplicações capaz de analisar grandes volumes de dados. A IA já faz parte do nosso cotidiano, desde escolher as melhores rotas em aplicativos de trânsito até cuidar da nossa saúde (FIGUEROA, 2020). De acordo com (FIGUEROA, 2020) a medicina é uma das áreas que mais tem se beneficiado da IA.

De acordo com a Organização Mundial da Saúde (OMS), a inteligência artificial (IA) é uma grande promessa para melhorar a prestação de atenção à saúde e medicamentos em todo o mundo, mas apenas se a ética e os direitos humanos forem colocados no centro de seu desenho, implantação e uso, de acordo com as novas orientações da organização (OMS, 2021).

O relatório "Ethics and governance of artificial intelligence for health"(Ética e governança da inteligência artificial para a saúde, em tradução livre) é resultado de dois anos de consultas realizadas por um painel de especialistas internacionais indicados pela OMS (OPAS, OMS, 2021).

Ao utilizar a Inteligência Artificial de forma responsável e ética na área da saúde, é possível obter diversas contribuições, como melhoria da eficiência no atendimento aos pacientes e na prevenção de doenças. Dentre as técnicas do Aprendizado de Máquina utilizadas no tratamento de doenças, destacam-se os algoritmos de regressão (OPAS, OMS, 2021).

3.3.1 Machine Learning

Em 1959, Arthur Samuel, engenheiro do MIT e considerado o precursor da inteligência artificial, abordou pela primeira vez o termo Machine Learning (ML). Ele também é creditado como criador do termo "Machine Learning"naquele mesmo ano, descrevendo o conceito como "um campo de estudo que dá aos computadores a habilidade de aprender sem terem sido programados para realização de tal tarefa"(QUARESMA, 2019).

De acordo com (GOMES, 2019), o aprendizado de máquina é o “campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados”.

Em 2006. Geoffrey et al. publicaram um artigo que descrevia como treinar uma rede neural profunda capaz de reconhecer dígitos manuscritos com uma precisão superior a 98%, o que foi considerado estado-da-arte na época. Essa técnica, denominada "Deep Learning"ou "Aprendizado Profundo", é um ramo específico do Aprendizado de Máquina (GÉRON, 2017).

De acordo com (GRUS, 2019), o aprendizado de máquina refere-se à criação e uso de modelos que são aprendidos a partir de dados, e pode ser denominado como modelo preditivo ou mineração de dados em outras definições. Para a presente pesquisa, será adotada a definição de aprendizado de máquina proposta pelo autor.

Para (GRUS, 2019) a árvore de decisão como um exemplo de como prever um resultado e destaca sua importância para o desenvolvimento do trabalho. Ele menciona que a utilização da árvore de decisão é uma boa maneira de identificar as doenças dos pacientes e, posteriormente, usar os dados para encontrar a melhor árvore de decisão para o problema em questão.

O Machine Learning, como subárea de pesquisa da Inteligência Artificial, tem como objetivo desenvolver técnicas computacionais e sistemas capazes de aprender conhecimento automaticamente (MONARD; BARANAUSKAS, 2003). Para isso, algoritmos são utilizados para que uma máquina tome decisões com base em dados prévios e nos dados utilizados pelo usuário, tornando possível a resolução de problemas complexos (MILDENBERGER, 2014).

A aplicação das técnicas de IA tem impactado diversas áreas, incluindo a saúde, onde os modelos de ML podem ser utilizados para prevenção e detecção precoce de doenças, além de evitar intervenções desnecessárias. Por exemplo, é possível selecionar grupos de pacientes para aplicação de vacinas ou prevenção da instalação de comorbidades em pacientes internados, o uso de ML na saúde tem potencial para melhorar a qualidade de vida da população, oferecendo diagnósticos mais precisos e tratamentos personalizados (MILDENBERGER, 2014).

Na busca pelo melhor modelo de ML para uso na área da saúde, é fundamental realizar uma revisão da literatura a fim de identificar trabalhos já desenvolvidos sobre o tema. Um

exemplo de trabalho que se destaca nesse sentido é o de (WENG et al., 2017), que concluiu que o uso de técnicas de aprendizado de máquina pode significativamente melhorar a precisão na previsão de RDC, aumentando o número de pacientes identificados que podem se beneficiar de tratamentos preventivos, e evitando outros tratamentos desnecessários. É importante ressaltar que o modelo de redes neurais apresentou o melhor desempenho dentre os modelos analisados, resultando em 355 predições corretas adicionais de doença cardiovascular quando comparado ao modelo baseado nas recomendações do Colégio Americano de Cardiologia. Com isso, pode-se afirmar que o uso de ML na saúde tem o potencial de melhorar significativamente a qualidade dos diagnósticos e tratamentos, contribuindo para a promoção de uma assistência médica mais eficiente e efetiva (WENG et al., 2017).

Os algoritmos de ML utilizados neste estudo foram modelos de regressão, que pertencem ao aprendizado supervisionado. Para caracterizar uma regressão, é necessário um conjunto de dados com n observações, cada uma representada por um vetor de mensurações das m variáveis independentes (ou variáveis preditoras), bem como a mensuração correspondente da variável dependente (ou variável resposta). Se a variável resposta for uma variável quantitativa, ela se enquadra nesse contexto. Além disso, as tarefas de ML podem ser divididas em duas categorias: preditivas e descritivas. No aprendizado supervisionado, a meta é encontrar uma função a partir dos dados de treinamento que possa ser utilizada para prever um rótulo ou valor que caracterize um novo exemplo, com base nos valores de seus atributos de entrada. Os algoritmos utilizados nessa tarefa são o K-Vizinhos-Mais-Próximos, Regressão Linear, Regressão Logística, Árvore de Decisão e Floresta Aleatória, Máquinas de Vetores de Suporte e Decision Tree Classification (GÉRON, 2017). (FACELI et al., 2011), também apresentam uma abordagem de aprendizado de máquina, abordando diversas técnicas, incluindo algoritmos supervisionados de regressão, como uma das possibilidades para a análise de dados.

Considerando a relevância e efetividade dos modelos preditivos de Machine Learning, é crescente o interesse e a adoção destas técnicas em diversas áreas do conhecimento, incluindo a Saúde. A literatura tem demonstrado a utilidade desses modelos na previsão de diagnósticos, na identificação de riscos, no suporte à decisão clínica e em outras aplicações importantes.

4 RELATO DO PROBLEMA

De acordo com a primeira Diretriz Brasileira de Diagnóstico e Tratamento da SM, essa condição se tornou um dos principais desafios na prática clínica atual. A presença da SM associada a DCV aumenta a taxa de mortalidade geral em cerca de 1,5 vezes e a taxa de mortalidade cardiovascular em aproximadamente 2,5 vezes, conforme destacado por (BRANDÃO, 2005).

Neste estudo, iniciamos nossa investigação analisando a possibilidade de desenvolvimento de risco de doença cardiovascular nos próximos 10 anos. O conjunto de dados para análise foi extraído da iniciativa NHANES [1]. O experimento empregou os critérios do FRS e alguns marcadores utilizados no diagnóstico da presença da SM, com exceção da medida do volume da cintura.

Os resultados revelaram uma diferença significativa entre a média do Risco de Evento Cardiovascular (RiscoECV) na população com SM (6.19) e a média na população sem SM (3.10). Para avaliar a magnitude dessa diferença, calculamos a razão entre essas médias.

Observamos que a média do Risco de Evento Cardiovascular (RiscoECV) na população com SM é quase o dobro da média na população sem SM, sendo aproximadamente 1,99 vezes maior. É importante ressaltar que, de acordo com (BRANDÃO, 2005), nenhum escore de risco, incluindo o estudo de Framingham, foi desenvolvido para prever a ocorrência futura de insuficiência cardíaca (IC) em pacientes com diabetes mellitus (DM) e SM.

Além disso, é relevante destacar que, até a atualização da Diretriz de Prevenção Cardiovascular da Sociedade Brasileira de Cardiologia em 2019, nenhum escore de risco específico para IC futura havia sido desenvolvido nesse grupo de pacientes, conforme mencionado por (PRECOMA, 2019). Portanto, considerando o objetivo deste trabalho, que é propor um Modelo de Aprendizado de Máquina para a Predição de RDC Utilizando Biomarcadores da SM e do FRS.

A obesidade é amplamente reconhecida como um fator de risco significativo para o desenvolvimento da SM. Segundo especialistas, como a endocrinologista Kristhiane Di Domenico, embora a predisposição genética possa influenciar a suscetibilidade à obesidade, hábitos de vida não saudáveis desempenham um papel fundamental no favorecimento do excesso de peso e da obesidade. Esses hábitos, como uma alimentação inadequada e a falta de atividade física regular, podem levar ao acúmulo de gordura corporal, desencadeando uma série de alterações metabólicas e fisiológicas (DOMENICO, 2021).

[1] disponível em: https://data.world/

A obesidade, por sua vez, contribui para o surgimento de outras condições que compõem a SM, como alterações na glicemia, no perfil lipídico e na pressão arterial. Dessa forma, a obesidade é considerada a base da síndrome, pois seu impacto negativo no organismo pode desencadear uma cascata de eventos que resultam em complicações metabólicas e cardiovasculares (DOMENICO, 2021).

A obesidade central é um dos principais marcadores utilizados para o diagnóstico da SM. Vale destacar que a obesidade é um problema global e sua prevalência vem aumentando de forma preocupante em todo o mundo. No Brasil, segundo dados de 2018 divulgados pelo Ministério da Saúde, 19,8% da população é obesa e mais da metade (55,7%) apresenta sobrepeso. Diante desses dados alarmantes, é fundamental que a obesidade seja tratada com atenção especial. No caso da SM, em particular, a obesidade central é um fator ainda mais preocupante que a obesidade total (DOMENICO, 2021).

A obesidade central não se refere apenas ao excesso de peso em si, mas sim ao aumento da circunferência da cintura. Em outras palavras, quando falamos de obesidade central, não estamos nos referindo ao Índice de Massa Corporal (IMC), mas sim à medida da cintura. Um valor igual ou superior a 88 cm para mulheres e 102 cm para homens é um sinal de alerta. Portanto, para reduzir a obesidade central, simplesmente fazer dieta pode não ser suficiente. É importante combinar uma alimentação adequada com exercícios físicos. Essa abordagem deve ser orientada aos pacientes diagnosticados com SM (DOMENICO, 2021).

Dentre as doenças cardíacas que podem ser causadas pela SM, destaca-se a doença arterial coronariana, causada pela obstrução das artérias coronárias, e a insuficiência cardíaca, que ocorre quando o coração não consegue bombear sangue suficiente para o corpo. Além disso, a SM também aumenta o risco de arritmias cardíacas, como a fibrilação atrial (American Heart Association, 2020) (ASSOCIATION, 2020).

De acordo com Grundy (GRUNDY, 2016), além da doença arterial coronariana e da insuficiência cardíaca, a síndrome metabólica pode estar relacionada a outras condições que aumentam o risco de doenças cardíacas, como a apneia do sono, a esteatose hepática não alcoólica e o diabetes tipo 2. Por isso, é fundamental que o tratamento da síndrome metabólica inclua o controle desses fatores de risco, a fim de prevenir o desenvolvimento de complicações cardiovasculares e outras consequências associadas à condição.

Diante da associação entre a SM e o RDC, mencionada por (BRANDÃO, 2005), tornou-se necessário elaborar uma Diretriz Brasileira de Diagnóstico e Tratamento da SM. O objetivo dessa diretriz foi reunir as principais evidências científicas disponíveis e apresentar recomendações

para o diagnóstico, tratamento e acompanhamento dessa condição no contexto brasileiro. No entanto, é importante ressaltar que, embora existam diversos escores de RDC na literatura, nenhum deles foi desenvolvido especificamente para a população com SM, o que limita sua aplicação clínica e sua capacidade de prever eventos cardiovasculares nessa população.

Além disso, com base na atualização de 2019 da Diretriz de Prevenção Cardiovascular da Sociedade Brasileira de Cardiologia, constatou-se que até o momento não foi desenvolvido nenhum escore de risco específico para a previsão de insuficiência cardíaca em pacientes com diabetes mellitus e SM (Precoma et al., 2019).

Considerando o levantamento realizado na revisão bibliográfica e a elaboração de um protocolo de revisão sistemática da literatura (RSL) para complementar e embasar este trabalho, buscou-se identificar no estado da arte se havia sido desenvolvido um escore que pudesse ser utilizado no diagnóstico e predição da SM e do RDC. Uma vez que a SM é um fator predominante no desenvolvimento desse risco, a revisão da literatura destacou a falta de um escore de RDC específico para a população com SM e DM. Diante dessa lacuna, surge a seguinte pergunta de pesquisa: "Qual a recomendação de um modelo de Machine Learning para a predição e diagnóstico da SM e RDC, explorando biomarcadores de diagnóstico da SM e o escore de risco de Framingham (FRS)?".

A proposta desse modelo é fornecer uma ferramenta útil para orientar a tomada de decisão clínica e melhorar a prevenção e o tratamento de DCV em pacientes com SM.

4.1 TRABALHOS RELACIONADOS

Esta seção apresenta uma revisão dos trabalhos nos quais os pesquisadores desenvolveram sistemas utilizando algoritmos de aprendizado de máquina. Além disso, descreve as metodologias empregadas para validar os modelos propostos pelos pesquisadores.

Ao longo desses estudos, diversos métodos de validação foram utilizados para avaliar a eficácia e a precisão dos modelos de aprendizado de máquina desenvolvidos. Entre as metodologias empregadas, destaca-se a validação cruzada, a divisão do conjunto de dados em conjuntos de treinamento e teste, a utilização de métricas de desempenho, como acurácia, precisão, recall e F1-score, e a comparação dos resultados com outros modelos ou abordagens existentes.

Além disso, muitos pesquisadores adotaram uma abordagem rigorosa ao realizar experimentos controlados, garantindo que seus modelos fossem testados em conjuntos de dados representativos e variados, abrangendo diferentes populações e características clínicas. Essa abordagem permitiu a avaliação da generalização dos modelos e a análise da sua aplicabilidade

em diferentes cenários.

Vale ressaltar que a validação dos modelos de aprendizado de máquina é um processo contínuo e interativo, no qual os pesquisadores buscam aprimorar suas abordagens e enfrentar desafios específicos relacionados aos dados disponíveis, à qualidade das variáveis e à escolha adequada dos algoritmos. Portanto, a utilização de metodologias robustas de validação é fundamental para garantir a confiabilidade e a aplicabilidade dos modelos propostos.

Ao revisar esses estudos, foi possível observar uma diversidade de abordagens e metodologias adotadas, cada uma com suas vantagens e limitações. Essa variedade reflete a complexidade e a multidimensionalidade dos problemas abordados e ressalta a importância de considerar cuidadosamente a escolha das metodologias de validação, considerando as peculiaridades do domínio de aplicação e dos conjuntos de dados utilizados.

A revisão desses trabalhos e metodologias de validação será fundamental para embasar a escolha da abordagem mais adequada na construção e validação do modelo de aprendizado de máquina proposto.

4.1.1 Revisão Bibliográfica e Revisão Sistemática da Literatura

Para a seleção dos estudos relacionados a esta dissertação, foi conduzida uma revisão bibliográfica abrangente com o objetivo de analisar o estado da arte no campo da predição de RDC relacionado à SM. Durante essa análise, constatou-se uma extensa quantidade de estudos que abordavam a predição de RDC na população em geral. No entanto, identificou-se uma lacuna de estudos específicos que investigassem o RDC na população com SM, utilizando técnicas de aprendizado de máquina.

Diante dessa lacuna, tornou-se evidente a importância de desenvolver um Revisão Sistemática da Literatura para auxiliar na seleção criteriosa de trabalhos relevantes. Por meio desse protocolo, foram selecionados minuciosamente três estudos que satisfazem os critérios estabelecidos. O objetivo principal desta pesquisa é propor uma solução eficaz e recomendar um modelo adequado para preencher essa lacuna de pesquisa, beneficiando não apenas a comunidade acadêmica, mas também áreas correlatas.

Conforme descrito na FIGURA 1, a coleta de dados foi conduzida seguindo a estratégia de RSL proposta por (KITCHENHAM, 2007). Nessa etapa, foram adotados cinco passos essenciais: busca e seleção dos estudos, avaliação da qualidade, extração dos dados, síntese dos resultados e interpretação dos achados. A ferramenta Parsifal, foi utilizada para a elaboração do protocolo[2]. O

[2] Acesso em: <https://parsif.al/> [12 de junho de 2021]

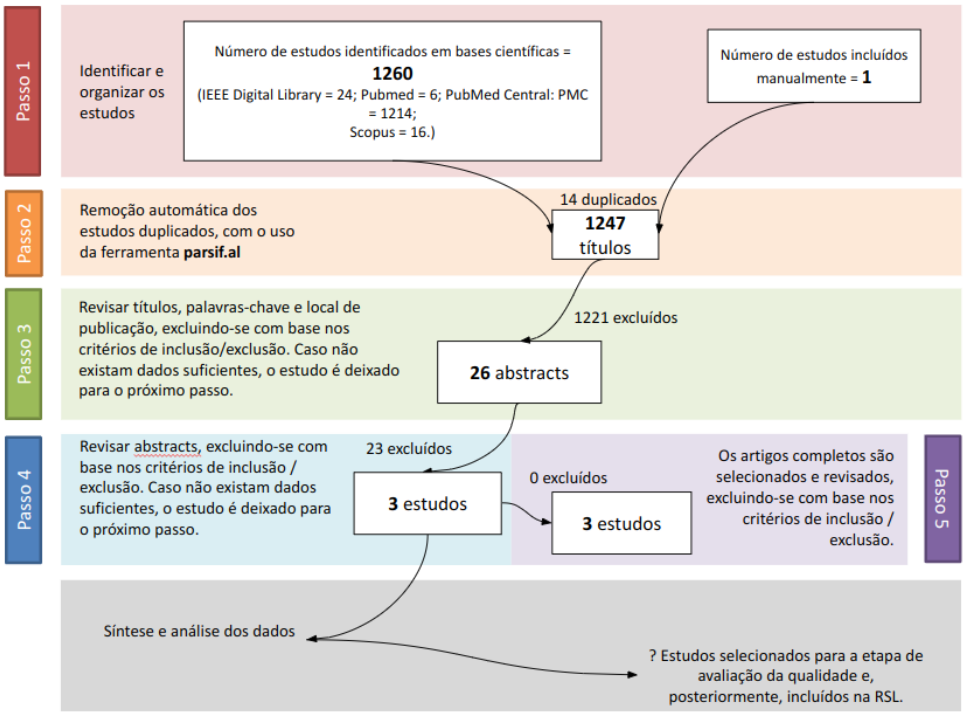
Parsifal é uma ferramenta online projetada para auxiliar pesquisadores na realização de revisões sistemáticas da literatura no contexto da Engenharia de Software.

No que se refere ao passo de seleção dos estudos, foram estabelecidos os seguintes critérios para a escolha das bases científicas: (1) inclusão de bases indexadas no Portal de Periódicos CAPES e (2) relevância das bases para as áreas de Aprendizado de Máquina, Modelos de Predição, SM, Risco Vascular e Escore de Risco Global de Framingham. Com base nesses critérios, as bases científicas selecionadas foram: IEEE Digital Library, PubMed, PubMed Central (PMC) e Scopus. Essa seleção abrange estudos de alta relevância nessas áreas, garantindo uma abrangência adequada para a pesquisa. A quantidade de estudos selecionados está disponível na FIGURA 1.

Com base nas palavras-chave e sinônimos utilizados na elaboração do protocolo, e após diversas validações para contemplar os três estudos inicialmente planejados, foi elaborada a seguinte string de busca automática: (("Machine learning"AND "Síndrome Metabólica") AND ("RDC")). Essa string de busca foi construída para garantir a inclusão dos termos pertinentes e otimizar a precisão na seleção dos estudos.

Como trabalho futuro, pretende-se aprimorar esta RSL inicial e publicá-la. Além disso, planeja-se incorporar todos os passos necessários para finalizar a RSL, a fim de obter uma visão abrangente do tema e identificar possíveis lacunas adicionais.

Figura 1 – Revisão Sistemática da Literatura



Fonte: : Autor desta dissertação

4.1.1.1 Estudos Selecionados

(PINILLA, 2021) informa que na Colômbia existem deficiências no desenvolvimento e desenho de modelos de previsão da SM usando técnicas de IA, portanto, sua proposta, foi desenvolver um sistema para prever a ocorrência de SM nos dados obtidos no estudo transversal e descritivo da FUPRECOL (Fuerza Prensil Colombia) utilizando técnicas de IA de alto desempenho como Máquinas de Vetores de Suporte e Redes Neurais, além de classificadores básicos. como Reconhecedor Euclidiano.

Para o processamento dos dados (PINILLA, 2021), realizou com a criação de padrões aleatórios utilizando a técnica ADASYN - para conjuntos de datos desequilibrados) e foram utilizadas técnicas de redução de dimensionalidade como FSCNCA - e Sequential Forward Selection. Os melhores modelos de cada técnica foram escolhidos e avaliados com padrões de teste para validar a porcentagem real da classificação do sistema.

As porcentagens de classificação foram obtidas com Redes Neurais e Máquinas de Vetor de Suporte de 91,8% e 93% respectivamente, sendo esta última a técnica com melhor predição. Os resultados obtidos mostram ser possível desenvolver sistemas com técnicas de IA para prever a SM com altas porcentagens de classificação, o que poderia ter grande potencial para prever desde a juventude a probabilidade que uma pessoa possa sofrer dessa síndrome e assim prevenir em idades adultas o aparecimento de doenças relacionadas a SM (PINILLA, 2021).

(YANG, 2018), no artigo "*Comparison between Metabolic Syndrome and the Framingham Risk Score as Predictors of Cardiovascular Diseases among Kazakhs in Xinjiang,*"os autores avaliou a eficácia da SM e do FRS como preditores de DCV em uma população cazaque em Xinjiang, China. Os autores coletou dados de 2.286 participantes e os acompanharam por um período médio de 5,49 anos. Eles descobriram que tanto a SM quanto o FRS foram capazes de prever o risco de DCV em indivíduos cazaques, mas que a SM pode ser um indicador mais efetivo de RDC em comparação com o FRS. Os resultados sugerem que a detecção precoce da SM pode ajudar a identificar indivíduos com maior risco de DCV e que medidas preventivas e de gerenciamento devem ser tomadas para reduzir o risco de DCV. Os autores também mostraram que a adição da idade ao escore de risco de SM aumentou significativamente sua sensibilidade e AUC, tornando-o um melhor preditor de DCV do que o FRS para a população estudada.

O estudo realizado por (MAIGA, 2019), comparou vários modelos de aprendizado de máquina para prever DCV usando um conjunto de dados da plataforma Kaggle. O conjunto de dados inclui 70.000 registros de pacientes com 12 atributos, incluindo idade, altura, peso, sexo, pressão arterial sistólica e diastólica, colesterol, glicose, tabagismo, consumo de álcool, atividade

física e presença ou ausência de DCVs. O objetivo do estudo foi determinar o melhor modelo para prever a presença de doença cardiovascular. Os modelos testados foram Random Forest, Naïve Bayes, KNN e Regressão Logística.

Os resultados do estudo mostraram que o modelo de floresta aleatória teve a melhor precisão na classificação de 73%, especificidade de 65% e sensibilidade de 80% na previsão de DCV. O autor do estudo enfatizou a importância da seleção de recursos para melhorar a precisão dos modelos de previsão de DCV.

Os resultados indicaram que o modelo de floresta aleatória teve a melhor precisão na previsão da presença de doença cardiovascular. Os resultados deste estudo são relevantes para a identificação precoce de DCV e podem ajudar a melhorar os esforços de prevenção e tratamento.

(SUDHA et al., 2021) comparou a precisão de vários métodos de aprendizado de máquina para prever o risco de DCV. O objetivo do estudo foi criar um modelo de previsão para auxiliar as pessoas a conhecerem o estado de saúde do coração. Para isso, foram coletados dados do escore de risco de Framingham padrão (FRS) e das características derivadas do ECG, e a técnica de fusão de recursos foi utilizada para combinar esses conjuntos de dados em um único banco de dados.

O estudo seguiu as seguintes etapas: coleta dos parâmetros FRS e ECG, fusão de dados no conjunto de dados FRS e ECG, pré-processamento de dados, construção de modelos de aprendizado de máquina, avaliação do desempenho dos modelos e previsão da pontuação de risco.

Os resultados do estudo indicaram que, individualmente, os principais fatores do conjunto de dados FRS e os parâmetros de ECG produziram precisões variando de 64,47% a 83,18%, dependendo do modelo de aprendizado de máquina utilizado. No entanto, essas medidas por si só não conseguiram explicar todos os fatores de risco que os pacientes com DCV apresentam. Como resultado, o procedimento de fusão de características do conjunto de dados de ECG foi usado para cobrir a maioria dos fatores de risco que levam ao desenvolvimento de uma pontuação de risco de DCV em indivíduos.

Em resumo, este estudo comparou vários modelos de aprendizado de máquina na previsão do risco de DCV, utilizando dados de escore de risco de Framingham e características derivadas do ECG. O estudo destacou a importância da fusão de características para melhorar a precisão dos modelos de previsão de DCV e contribui para a identificação precoce e tratamento dessas doenças.

De acordo com (HUANG et al., 2021) o artigo *Application of ensemble machine learning*

algorithms on lifestyle factors and wearables for cardiovascular risk prediction aborda fontes de dados inovadoras para a previsão do RDC, incluindo questionários detalhados sobre estilo de vida e monitoramento contínuo da pressão arterial, utilizando MLAs. O escore de risco convencional de referência utilizado foi o FRS.

Os resultados obtidos revelaram que os MLAs em conjunto, construídos com base nos algoritmos Naive Bayes, Random Forest e Support Vector Classifier para categorias de baixo risco, e Generalized Linear Regression, Support Vector Regressor e Stochastic Gradient Descent Regressor para categorias de alto risco, superaram o desempenho do FRS tanto para as categorias de baixo quanto de alto risco. O MLA baseado apenas no questionário sobre estilo de vida obteve uma AUC de 0,715 (95% CI 0,681, 0,750) e 0,710 (95% CI 0,653, 0,766) para baixo e alto risco, respectivamente.

Ao combinar todos os grupos de fatores de risco, como questionários sobre estilo de vida, exames clínicos de sangue, monitoramento de pressão arterial ambulatorial de 24 horas e frequência cardíaca, juntamente com a seleção de características, a previsão dos grupos de baixo e alto risco de DCV foi aprimorada ainda mais, alcançando uma AUC de 0,791 (95% CI 0,759, 0,822) e 0,790 (95% CI 0,745, 0,836) respectivamente.

Além dos preditores convencionais, a atividade física autorrelatada, a frequência cardíaca diária média, a variabilidade da pressão arterial ao acordar e a porcentagem de tempo em hipertensão diastólica foram identificados como importantes contribuintes para a classificação de RDC.

Portanto, este estudo oferece percepções valiosos sobre a utilização de fontes de dados inovadoras e algoritmos de aprendizado de máquina em conjunto para a previsão do RDC. Os resultados ressaltam a importância de uma abordagem abrangente e integrada que incorpora diferentes aspectos do estilo de vida e parâmetros fisiológicos na avaliação do RDC. Essas descobertas têm implicações significativas no desenvolvimento de estratégias personalizadas e mais precisas na prevenção e tratamento de DCV.

É importante salientar que os conjuntos de dados que suportam os resultados deste estudo não estão disponíveis publicamente devido à proteção de dados pessoais e razões éticas.

5 PROPOSTA

Neste capítulo, descrevemos a proposta do trabalho, incluindo os detalhes relevantes para a elaboração do modelo de aprendizado de máquina e as correlações das variáveis utilizadas.

5.1 FUNDAMENTAÇÃO

Com o avanço da inteligência artificial e do aprendizado de máquina, tornou-se possível criar modelos de previsão e diagnóstico do risco cardíaco relacionado à SM com maior precisão e eficiência. Uma proposta viável é utilizar os marcadores da SM e os dados do estudo Framingham para selecionar o melhor modelo de aprendizado de máquina.

Os marcadores da SM abrangem diversos aspectos, como obesidade abdominal, níveis elevados de triglicerídeos, baixos níveis de HDL, pressão arterial elevada e glicose elevada em jejum. Para fundamentar este trabalho, recorremos ao estudo Framingham, um estudo longitudinal reconhecido por fornecer dados confiáveis sobre os fatores de RDC, tais como idade, sexo, índice de massa corporal (IMC), tabagismo, colesterol total, HDL-colesterol, pressão arterial, entre outros.

Com base nas informações e indicadores apresentados, este estudo empregará algoritmos de aprendizado de máquina. Modelos, tais como Regressão Logística, SVR, KNeighborsRegressor, Árvore de Decisão, Gradient Boosting Regressor e Random Forest Regressor, serão considerados para a construção de um modelo de previsão do risco cardíaco associado à SM.

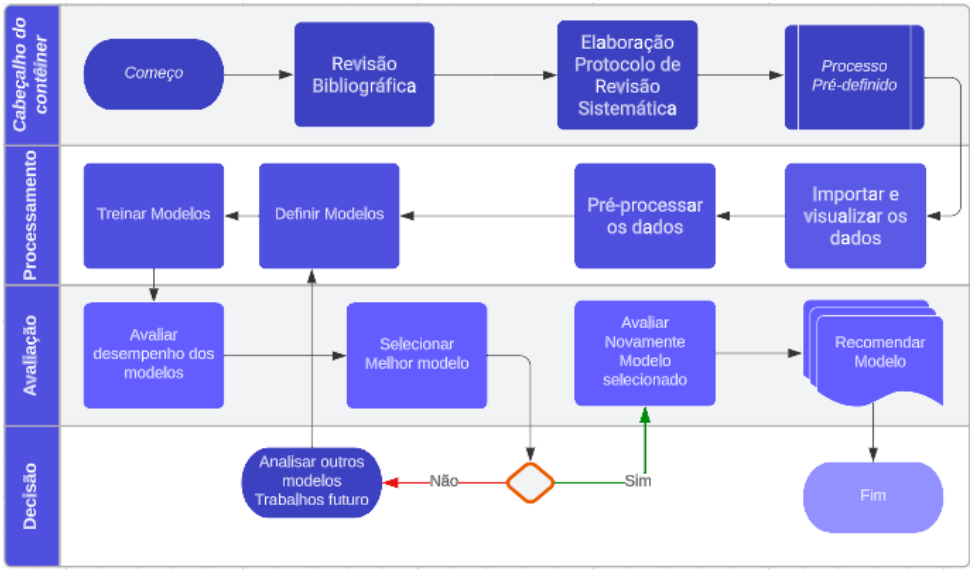
Nos cenários 1, 2, 3 e 4, o modelo será treinado utilizando conjuntos de dados de pacientes que apresentam esses fatores de risco. No que concerne aos experimentos realizados no cenário 5, o conjunto de dados utilizado compreende a combinação de quatro bancos de dados e concentra-se primordialmente na diferenciação da presença de cardiopatias.

No sexto e último cenário, faremos uso de uma base de dados que não apenas abrange alguns dos fatores de risco previamente mencionados, mas também registra eventos cardiovasculares, tais como infarto ou acidente vascular cerebral, conhecido como o conjunto de Framingham.

Essa abordagem permitirá a identificação e avaliação do risco cardíaco em pacientes com SM, fornecendo subsídios para a prevenção e o tratamento adequado dessa condição. A integração dos fatores de RDC e eventos cardiovasculares na análise enriquecerá a capacidade do modelo em oferecer insights valiosos para a prática clínica e o desenvolvimento de estratégias de intervenção mais eficazes.

Após o treinamento do modelo, ele pode ser aplicado para prever o RDC em pacientes com marcadores associados à SM. Isso permite que os médicos adotem medidas preventivas para reduzir esse risco. Além disso, o modelo pode auxiliar no diagnóstico precoce, identificando os pacientes com maior probabilidade de desenvolver DCV e, consequentemente, necessitando de um cuidado mais intensivo. Portanto, a utilização de algoritmos de aprendizado de máquina em conjunto com os marcadores da SM e os dados do estudo Framingham oferece uma solução eficiente e precisa para a prevenção e diagnóstico do risco cardíaco relacionado à SM.

Figura 2 – Fluxograma aplicado para recomendação do modelo



Fonte: : Elaboração própria (2023)

5.2 HIPÓTESE

A hipótese central deste estudo sustenta que os bioindicadores, que incluem as variáveis relacionadas à SM de acordo com o critério NCEP-ATP III, e os marcadores do FRS, exercem um impacto positivo, forte e significativo nos eventos cardiovasculares em pacientes diagnosticados com SM.

Além disso, considera-se que esses bioindicadores refletem características e condições que estão diretamente associadas ao desenvolvimento e progressão das DCV em pacientes com SM. Acredita-se que a presença de componentes da SM, como obesidade abdominal, hipertensão arterial, dislipidemia e resistência à insulina, juntamente com os fatores de risco identificados pelo FRS, como idade, sexo, tabagismo e níveis elevados de colesterol, contribuam para um ambiente propício ao surgimento de eventos cardiovasculares.

5.3 RELEVÂNCIA

A SM é uma condição complexa e multifatorial que está associada a um aumento do RCD e outras complicações de saúde. A identificação precoce de pacientes com SM e o gerenciamento eficaz de seus fatores de risco são essenciais para prevenir DCV e melhorar a qualidade de vida. Nesse contexto, a utilização de modelos de ML pode ser uma ferramenta valiosa para prever o RDC em pacientes com SM e orientar os profissionais de saúde na escolha das melhores estratégias terapêuticas. Portanto, explorar a precisão dos modelos de ML na previsão de DCV em pacientes com SM é uma pesquisa relevante e atual, com o potencial de melhorar o cuidado com os pacientes e reduzir a incidência de DCV.

A utilização de modelos de ML oferece vantagens significativas na área da saúde, pois permite o processamento de grandes quantidades de dados de forma rápida e precisa. Ao combinar os marcadores da SM e os marcadore de FRS, nosso estudo busca desenvolver um modelo de previsão que considere a complexidade e as interações entre os fatores de RDC. Acreditamos que a utilização desses modelos contribuirá para uma abordagem mais personalizada no cuidado dos pacientes com SM, permitindo a identificação dos indivíduos com maior risco de desenvolver DCV e a implementação de intervenções precoces.

Além disso, a aplicação de modelos de ML na previsão de DCV em pacientes com SM tem implicações clínicas significativas. Ao identificar de forma precisa os pacientes com maior risco, os profissionais de saúde podem adotar medidas preventivas e personalizadas, como a prescrição de medicamentos adequados, a implementação de mudanças no estilo de vida e a adoção de estratégias de acompanhamento mais intensivas. Isso não apenas reduzirá a incidência de DCV, mas também melhorará a qualidade de vida dos pacientes, reduzindo hospitalizações e complicações relacionadas a essas condições.

Portanto, a pesquisa proposta tem relevância tanto para a comunidade científica quanto para a prática clínica. A validação das hipóteses e a obtenção de resultados confiáveis e precisos no desenvolvimento do modelo de previsão fortalecerão a base de evidências e fornecerão subsídios para a implementação de estratégias de prevenção e tratamento mais eficazes em pacientes com SM. Isso terá um impacto positivo na saúde pública, ajudando a reduzir a carga das DCV e melhorar a qualidade de vida das pessoas afetadas pela SM.

6 METODOLOGIA DE EXPERIMENTO

Neste capítulo, será detalhado o processo de desenvolvimento do modelo, incluindo as etapas envolvidas e as abordagens utilizadas. Serão exploradas técnicas avançadas de análise de dados e aprendizado de máquina, com o objetivo de identificar padrões e relacionamentos relevantes para a predição e classificação do RDC.

Além disso, serão descritos os instrumentos e dispositivos que serão utilizados durante a coleta de dados. Serão adotados métodos precisos e confiáveis de medição, levando em consideração os parâmetros-chave relacionados à SM e ao RDC.

A utilização dessa metodologia proporcionará uma base sólida para o desenvolvimento de um modelo preciso, capaz de auxiliar na identificação precoce e no gerenciamento do RDC em pacientes com SM.

6.1 MÉTODOS

Durante a busca por dados reais para utilização, foram encontrados três datasets relevantes. O primeiro deles é o "mswithinsulin2", que consiste em 29 colunas e 1.943 linhas. Esse conjunto de dados contém informações de 693 pacientes diagnosticados com SM e 1.250 pacientes sem essa condição clínica. Do total, 993 pacientes são do sexo feminino e 950 são do sexo masculino. A Tabela 1 apresenta os marcadores clínicos disponíveis nesse conjunto de dados. O autor responsável por este conjunto de dados é Robert Hoyt, MD (HOYT, 2018).

Os dados utilizados para construir o modelo usado nos cenários 1, 2, 3 e 4 foram obtidos a partir da base de dados "MSWithInsulin.csv", [1]. Esses dados foram extraídos do *National Center for Health Statistics* (NHANES) e incluíram informações de pacientes diagnosticados com SM e pacientes sem a doença, em um período de dez anos.

[1] disponível na plataforma: https://data.world/

Tabela 1 – Marcadores Clínicos do Dataset mswithinsulin2

Nome do Marcador	Descrição
seqn	Número de sequência da amostra
age	Idade do paciente em anos
sex	Sexo do paciente (0 = masculino, 1 = feminino)
marital_status	Estado civil do paciente
annual_income	Renda anual do paciente
race	Raça do paciente
fname	Nome do paciente
lname	Sobrenome do paciente
WaistCirc	Circunferência da cintura em cm
BMI	Índice de massa corporal do paciente
albuminuria	Nível de albuminúria na urina em mg/dL
UrAlbCr	Razão albumina/creatinina na urina
UricAcid	Nível de ácido úrico no sangue em mg/dL
GGT	Nível de gama-glutamiltransferase no sangue em U/L
ALT	Nível de alanina aminotransferase no sangue em U/L
AST	Nível de aspartato aminotransferase no sangue em U/L
CPK	Nível de creatina quinase no sangue em U/L
HOMA	Modelo homeostático de avaliação da resistência à insulina (HOMA-IR)
BloodGlucose	Nível de glicose no sangue em mg/dL
BloodInsulin	Nível de insulina no sangue em μU/mL
HDL	Nível de lipoproteína de alta densidade no sangue em mg/dL
Trigylcerides	Nível de triglicerídeos no sangue em mg/dL
smoking	Status do hábito de fumar do paciente
Hypertension	Status de hipertensão do paciente
Dyslipidemia_HDL	Status de dislipidemia de HDL do paciente
Dyslipidemia	Status de dislipidemia do paciente
Hyperglycemia	Status de hiperglicemia do paciente
Obesity	Status de obesidade do paciente
MetabolicSyndrome	Status de síndrome metabólica do paciente

Fonte: : Elaboração própria (2023)

Com o intuito de conduzir experimentos com uma base de dados alternativa, no cenário 5 da pesquisa, foi necessário buscar dados reais para a sua utilização. Nesse contexto, identificou-se um segundo conjunto de dados originalmente composto por 76 atributos. No entanto, conforme mencionado por (JANOSI, 1988), os experimentos previamente publicados têm se concentrado em um subconjunto específico de 14 atributos. Esse conjunto de dados, originário do banco de dados de Cleveland, continua a ser amplamente empregado por pesquisadores de aprendizado de máquina até os dias atuais.

Com uma amostra populacional composta por 1.190 indivíduos, dos quais 909 são do sexo feminino e 281 são do sexo masculino, o atributo "objetivo"nesse conjunto de dados está

relacionado à presença de cardiopatia nos pacientes, sendo representado por um valor inteiro que varia de 0 (ausência) a 4.

A Tabela 2 apresenta os dados clínicos disponíveis nesse conjunto de dados. Esses dados incluem informações relevantes, como idade do paciente, sexo, tipo de dor no peito, pressão arterial em repouso, níveis de colesterol e glicose em jejum, eletrocardiograma em repouso, frequência cardíaca máxima atingida, presença de angina induzida pelo exercício, depressão do segmento ST induzida pelo exercício, inclinação do segmento ST e a presença ou ausência de cardiopatia.

Tabela 2 – Marcadores Clínicos do Dataset Cleveland, Hungria, Suíça e VA Long Beach

Nome do Marcador	Descrição
age	Idade do paciente em anos
sex	Sexo do paciente (0 = masculino, 1 = feminino)
chest pain type	Tipo de dor no peito
resting bp s	Pressão arterial em repouso (mm Hg)
cholesterol	Colesterol (mg/dL)
fasting blood sugar	Nível de glicose em jejum (mg/dL)
resting ecg	Eletrocardiograma em repouso
max heart rate	Frequência cardíaca máxima atingida
exercise angina	Angina induzida pelo exercício (0 = não, 1 = sim)
oldpeak	Depressão do segmento ST induzida pelo exercício
ST slope	Inclinação do segmento ST
target	Presença de cardiopatia (0 = ausência, 1 = presença)

Fonte: : Elaboração própria (2023)

Os experimentos realizados com o banco de dados de Cleveland têm se focado principalmente na distinção entre a presença de cardiopatia (valores 1, 2, 3, 4) e a sua ausência (valor 0). A fim de preservar a privacidade dos pacientes, os nomes e números de segurança social foram recentemente removidos e substituídos por valores fictícios. No diretório do banco de dados, há um arquivo "processado"que contém os dados já tratados do banco de dados de Cleveland, enquanto os outros quatro arquivos estão disponíveis sem processamento (JANOSI, 1988).

Nos experimentos conduzidos no sexto cenário, empregamos o terceiro conjunto de dados do projeto, denominado Dataset Framingham. Esse conjunto de dados contém mais de 4.240 registros, dos quais 2.420 são do sexo feminino e 1.820 são do sexo masculino. Dentre os registros, 2.095 são fumantes e 2.145 são não fumantes. Esses dados estão organizados em 16 colunas e fornecem informações sobre 15 atributos relevantes. O arquivo "framingham.csv"se refere a esse conjunto de dados, que foi originalmente coletado no estudo de Framingham. Esse estudo é uma pesquisa epidemiológica renomada sobre DCV, realizada em Framingham, nos Estados Unidos, a partir de 1948, com o objetivo de identificar os principais fatores de risco associados a doenças cardíacas (FRAMINGHAM. . . , 2022).

Os dados do conjunto de dados do FRS englobam uma variedade de informações demográficas, clínicas e relacionadas ao estilo de vida dos participantes do estudo de Framingham. Essas informações abrangem uma ampla gama de variáveis, como idade, sexo, pressão arterial, níveis de colesterol, histórico familiar de DCV e tabagismo, entre outras. Essa diversidade de dados proporciona uma compreensão abrangente dos indivíduos envolvidos no estudo, permitindo uma análise mais completa e precisa dos fatores de risco e sua relação com as DCV. A tabela 3 apresenta os marcadores clínicos disponíveis no conjunto de dados do FRS juntamente com suas descrições correspondentes.

Tabela 3 – Marcadores Clínicos do Dataset Framingham

Nome do Marcador	Descrição
male	Sexo do paciente (0 = feminino, 1 = masculino)
age	Idade do paciente em anos
education	Nível de educação
currentSmoker	Fumante atual (0 = não, 1 = sim)
cigsPerDay	Quantidade de cigarros fumados por dia
BPMeds	Uso de medicamentos para pressão arterial (0 = não, 1 = sim)
prevalentStroke	Presença de histórico de acidente vascular cerebral (0 = não, 1 = sim)
prevalentHyp	Presença de hipertensão (0 = não, 1 = sim)
diabetes	Presença de diabetes (0 = não, 1 = sim)
totChol	Nível total de colesterol (mg/dL)
sysBP	Pressão arterial sistólica (mm Hg)
diaBP	Pressão arterial diastólica (mm Hg)
BMI	Índice de Massa Corporal
heartRate	Frequência cardíaca
glucose	Nível de glicose (mg/dL)
TenYearCHD	Presença de doença cardíaca em 10 anos (0 = ausência, 1 = presença)

Fonte: : Elaboração própria (2023)

Devido à sua importância e relevância, esse conjunto de dados é amplamente utilizado

em pesquisas e estudos relacionados a DCV. Ele permite a análise da relação entre os fatores de risco e o desenvolvimento dessas doenças, auxiliando na identificação de padrões, tendências e estabelecimento de modelos de previsão e diagnóstico relacionados a problemas cardíacos.

Ao utilizar o arquivo "framingham.csv"como parte do projeto em questão, é possível explorar e investigar a associação entre os fatores de RDCes e a SM, contribuindo para uma melhor compreensão e abordagem dessa condição médica específica.

6.2 ABORDAGEM PARA O DESENVOLVIMENTO E SELEÇÃO DO MODELO RECOMENDADO

Nesta seção, será exposta a metodologia adotada para o desenvolvimento e seleção do modelo de aprendizado de máquina mais apropriado para a predição do RDC associado à SM.

6.2.1 Ambiente de Implementação e Ferramentas Utilizadas:

Para a implementação dos algoritmos, foi utilizado um equipamento com uma CPU Intel Core i7-8700T, que possui 6 núcleos físicos e 12 threads. O sistema conta com uma memória de 16GiB. No que diz respeito ao sistema operacional, utilizamos o Ubuntu 22.04.2 LTS.

Para o desenvolvimento, a escolha foi pelo o ambiente de desenvolvimento Colab, uma plataforma gratuita fornecida pelo Google. Essa plataforma permite criar e executar códigos em Python diretamente no navegador, eliminando a necessidade de instalar softwares adicionais no computador. O Colab é baseado no Jupyter Notebook, o que facilita o compartilhamento e a colaboração em projetos. Além disso, o Colab oferece recursos avançados, como integração com serviços de nuvem do Google, incluindo o Google Drive e o Google Cloud. Também temos acesso gratuito a GPUs, o que acelera significativamente o processamento de dados, especialmente em tarefas de aprendizado de máquina e ciência de dados.

Com todas essas vantagens, o Colab se mostra uma excelente opção para trabalhar em projetos de aprendizado de máquina e ciência de dados, proporcionando um ambiente prático, colaborativo e com recursos poderosos.

Python foi a linguagem de programação escolhida devido à sua disponibilidade no ambiente Colab. É uma linguagem de alto nível, interpretada e orientada a objetos, amplamente utilizada em diversas áreas, incluindo ciência de dados, aprendizado de máquina, desenvolvimento web e automação de tarefas. A sintaxe clara e simples do Python, juntamente com sua vasta comunidade de desenvolvedores ativos e a disponibilidade de bibliotecas abrangentes, tornam essa linguagem uma das mais populares no mundo da programação. Sua curva de aprendizado é considerada amigável para iniciantes. Para obter mais informações, o site oficial [2] é uma excelente fonte de referência (PEDREGOSA, 2011).

As bibliotecas em Python são conjuntos de módulos que oferecem funções e ferramentas para realizar tarefas específicas de programação de forma mais rápida e eficiente. Entre as bibliotecas mais comuns, destacam-se pandas, numpy e scikit-learn, amplamente utilizadas em áreas como ciência de dados, aprendizado de máquina, processamento de imagens, entre outras.

[2] Acesso em: https://www.python.org/ [10 de junho de 2023]

Essas bibliotecas são desenvolvidas por membros ativos da comunidade Python e permitem que os programadores utilizem soluções já testadas e aprimoradas em seus projetos, reduzindo o tempo de desenvolvimento e aumentando a eficiência do código. Dessa forma, as bibliotecas em Python facilitam o trabalho dos desenvolvedores, permitindo a criação de soluções mais sofisticadas e precisas em áreas como inteligência artificial, análise de dados, entre outras (PEDREGOSA, 2011).

Passo 1: Treinamento e Avaliação de Modelos de ML:

Para a seleção do modelo, baseamos nossa escolha em modelos previamente empregados em estudos relacionados a esta pesquisa. Adicionalmente, em consonância com as diretrizes recomendadas em aprendizado de máquina, implementamos um processo estruturado de seleção e avaliação de modelos, que abrangeu uma série de etapas.

Durante a elaboração deste projeto, realizamos uma abrangente investigação de uma variedade de modelos de aprendizado de máquina com o intuito de identificar aquele que mais se alinhasse com os objetivos deste estudo. Para isso, implementamos um pipeline de seleção e avaliação de modelos. A seguir, apresentamos os modelos que foram submetidos a treinamento e avaliação:

Regressão Logística: um modelo de classificação que estima a probabilidade de uma variável de destino pertencer a uma determinada classe. É amplamente utilizado em problemas de classificação binária (GÉRON, 2017).

SVR: um modelo de regressão que utiliza vetores de suporte para estimar a função que melhor se ajusta aos dados. É especialmente eficaz em problemas com dados não lineares (GÉRON, 2017).

KNeighborsRegressor: um modelo de regressão baseado em vizinhos mais próximos, que estima o valor da variável de destino com base nos valores de suas instâncias vizinhas mais próximas (GÉRON, 2017).

Árvore de Decisão: um modelo que utiliza uma estrutura de árvore para tomar decisões com base em testes em diferentes atributos. É amplamente utilizado em problemas de classificação e regressão (GÉRON, 2017).

Gradient Boosting Regressor: um modelo de aprendizagem de máquina que constrói um conjunto de modelos fracos em sequência, onde cada modelo tenta corrigir os erros do modelo anterior. É conhecido por sua capacidade de lidar com dados complexos e fornecer previsões de alta precisão (GÉRON, 2017).

Random Forest Regressor: um modelo que combina várias árvores de decisão para

realizar a regressão. Cada árvore individual é treinada em uma amostra aleatória do conjunto de dados e, em seguida, suas previsões são combinadas para obter uma previsão final mais robusta (GÉRON, 2017).

Durante os testes desses modelos, conduzimos uma avaliação abrangente de métricas de desempenho, incluindo precisão, acurácia e erro médio quadrático, a fim de selecionar o que melhor se adequasse aos nossos requisitos. A escolha final do modelo foi fundamentada em sua capacidade de fornecer resultados precisos e eficazes na previsão do RDC associado à SM.

Além dos modelos de regressão mencionados anteriormente, também exploramos modelos de classificação durante o desenvolvimento do projeto. Abaixo estão alguns dos modelos de aprendizagem de máquina testados para classificar o RDC associado à SM:

Regressão Logística: como mencionado anteriormente, este modelo pode ser usado tanto para problemas de classificação binária quanto para regressão. Na classificação, ele estima a probabilidade de pertencer a uma determinada classe (GÉRON, 2017).

SVM: um modelo de classificação que mapeia os dados de entrada em um espaço dimensional mais alto e encontra o hiperplano que melhor separa as classes. É eficaz em problemas com dados não lineares e permite diferentes funções de kernel para lidar com diferentes tipos de dados (GÉRON, 2017).

KNN: um modelo de classificação que classifica novos pontos de dados com base nas classes dos vizinhos mais próximos. Ele calcula a distância entre os pontos e seleciona a classe majoritária dos vizinhos mais próximos (GÉRON, 2017).

Árvore de Decisão: como mencionado anteriormente, este modelo também é usado para problemas de classificação. Ele constrói uma árvore de decisão com base em testes em diferentes atributos e classifica os pontos de dados com base nas regras de divisão da árvore (GÉRON, 2017).

Gradient Boosting Classifier: semelhante ao Gradient Boosting Regressor, este modelo constrói um conjunto de classificadores fracos em sequência, onde cada classificador tenta corrigir os erros do classificador anterior. É conhecido por sua capacidade de lidar com dados complexos e fornecer previsões precisas (GÉRON, 2017).

Random Forest Classifier: similar ao Random Forest Regressor, este modelo combina várias árvores de decisão para realizar a classificação. Cada árvore individual é treinada em uma amostra aleatória do conjunto de dados e suas previsões são combinadas para obter uma previsão final mais robusta (GÉRON, 2017).

Durante a avaliação desses modelos, foram utilizdas métricas como acurácia, precisão,

recall e F1-score para selecionar aquele que melhor se ajustasse aos nossos requisitos. A escolha final do modelo de classificação foi baseada em sua capacidade de fornecer previsões precisas e confiáveis para a classificação do RDC em pacientes com SM.

Após a delimitação dos procedimentos preliminares para a realização dos experimentos no âmbito do projeto, esta subseção discorrerá sobre o ambiente de implementação e as ferramentas empregadas no contexto da importação, visualização e pré-processamento dos dados relacionados aos pacientes com SM e sem SM.

A escolha e configuração de todas as técnicas e modelos mencionados nesta seção foram fundamentadas na obra "Machine Learning in Python"(PEDREGOSA, 2011), que serviu como referência para a condução adequada dos estudos.

Passo 2: Importação e Visualização dos Dados

A etapa de Importação e Visualização dos Dados envolve a coleta dos dados relevantes de várias fontes, como arquivos CSV, bancos de dados ou APIs. Esses dados são carregados no ambiente de trabalho para serem processados e analisados. Em seguida, ocorre a exploração e visualização dos dados, utilizando gráficos, tabelas e outras técnicas visuais, a fim de compreender a estrutura, padrões e relacionamentos presentes nos dados. Essa análise visual permite identificar tendências, distribuições, outliers e insights importantes. Essa etapa é essencial para obter uma compreensão inicial dos dados, detectar possíveis problemas e direcionar as próximas etapas da análise. A importação e visualização dos dados fornecem uma base sólida para a tomada de decisões e o desenvolvimento de estratégias analíticas adequadas.

Passo 3: Pré-processamento dos Dados

O pré-processamento dos dados é uma etapa essencial na análise de dados, foram utilizadas várias técnicas para garantir a qualidade e a precisão dos resultados. Entre as técnicas aplicadas, destacam-se a limpeza de dados, a seleção de variáveis relevantes e a transformação dos dados, quando necessário. A limpeza de dados envolveu a remoção de dados duplicados, o preenchimento de valores ausentes e a exclusão de dados inconsistentes ou irrelevantes.

Além disso, foram selecionados os marcadores utilizados no estudo de Framingham, através de testes envolvendo tanto os marcadores de identificação da SM quanto os de Framingham. Essa seleção foi realizada visando prever o desenvolvimento de SM e DCV.

Passo 4: Treinamento de Diversos Modelos

Nesta etapa, o treinamento de diversos modelos de regressão foram realizados com o objetivo de identificar aquele que apresentasse o melhor desempenho na tarefa de previsão de DCV em pacientes com SM. Para isso, foi realizada a seleção e separação das features, ou seja,

das variáveis que seriam utilizadas na análise dos dados. Além disso, também foram treinados os algoritmos de classificação mencionados anteriormente nesta seção, aplicando-os para identificar a presença da doença em questão.

Pipiline foi a técnica de ML utilizada para realizar a avaliação dos diferentes modelos de aprendizado de máquina. Essa abordagem consiste na aplicação sequencial dos modelos em etapas consecutivas, o que permite uma comparação direta do desempenho entre eles. A fim de evitar o overfitting dos modelos aos dados de treinamento e garantir uma avaliação robusta, os dados foram divididos em conjuntos de treinamento e teste (PEDREGOSA, 2011).

No processo de treinamento dos modelos, foi empregada a técnica de validação cruzada com um número de folds igual a 5. A validação cruzada é uma estratégia utilizada para estimar o desempenho do modelo de forma mais precisa, onde o conjunto de dados é dividido em k partes iguais. Em cada iteração, k-1 partes são usadas como conjunto de treinamento e a parte restante é utilizada como conjunto de teste. Esse procedimento é repetido k vezes, garantindo que todas as partes do conjunto de dados sejam utilizadas tanto para treinamento quanto para teste. Dessa forma, a validação cruzada de 5 proporciona uma avaliação mais robusta e confiável dos modelos (PEDREGOSA, 2011).

Dessa forma, garantimos que os modelos sejam avaliados de forma imparcial e confiável, proporcionando resultados confiáveis para a previsão de DCV em pacientes com SM.

Passo 5: Avaliação do Desempenho dos Modelos utilizando Métricas de Classificação e Regressão

Na etapa de avaliação do desempenho dos modelos, foi utilizada as métricas de regressão, como o erro médio absoluto, o erro médio quadrático e o coeficiente de determinação. Empregamos a técnica de validação cruzada com 5 folds utilizando o GridSearchCV para buscar os melhores hiperparâmetros do modelo. Essa técnica foi importante para avaliar o desempenho do modelo de forma mais robusta e evitar problemas de overfitting e underfitting.

Passo 6: Avaliação Final do Modelo Selecionado

Por fim, realizou-se uma última avaliação do modelo selecionado para assegurar seu desempenho satisfatório na previsão do RDC em pacientes.

Ao longo deste estudo, além dos experimentos realizados no conjunto de dados mencionado no passo 2, também foram conduzidos experimentos em mais dois conjuntos de dados. Um deles é o conjunto de dados Framingham, que contém mais de 4.240 registros, 16 colunas e 15 atributos. Seu objetivo é prever o risco de CHD em pacientes para um período de 10 anos. Esse conjunto de dados está disponível no Kaggle (https://www.kaggle.com/datasets/aasheesh200/framingham-

heart-study-dataset). O outro conjunto de dados utilizado foi uma combinação de dados de doenças cardíacas de várias fontes, incluindo pacientes da Cleveland Clinic e Hungria, totalizando 1.190 pacientes. Mais informações sobre esse conjunto de dados podem ser encontradas no repositório de dados de Irvine da Universidade da Califórnia (https://archive-beta.ics.uci.edu/dataset/45/heart+disease) e nos anexos deste estudo (JANOSI, 1988).

Todos os algoritmos criados neste projeto serão disponibilizados nos apêndices deste trabalho.

6.3 EXPERIMENTOS

Na seção de Experimentos, foram realizadas uma série de avaliações com o intuito de analisar o desempenho e a eficácia dos modelos propostos. Diferentes conjuntos de dados foram utilizados, os quais foram importados e devidamente processados. Medidas de pré-processamento foram aplicadas para garantir a qualidade dos dados, incluindo técnicas de normalização e tratamento de valores ausentes. Em seguida, os modelos foram treinados e avaliados utilizando métricas apropriadas para problemas de classificação e regressão. Análises estatísticas e comparações entre os modelos foram conduzidas para identificar aqueles com melhor desempenho. Os resultados obtidos fornecem insights valiosos sobre a capacidade preditiva dos modelos e sua adequação para o problema em questão.

Além disso, nesta fase específica da pesquisa, o arquivo "MSWithInsulin.csv"foi utilizado como o conjunto de dados principal. Os dados foram importados e uma análise detalhada foi realizada para compreender a estrutura dos dados e identificar possíveis problemas que pudessem afetar os resultados. Durante esse processo, foram tomados cuidados especiais na inspeção e no processamento minucioso dos dados importados, garantindo sua integridade e qualidade.

6.3.1 Cenários

Após a normalização dos dados e a etapa de preparação para a implementação dos algoritmos, foi necessário selecionar os cenários mais adequados para atender às hipóteses e às propostas iniciais da pesquisa. Esses cenários foram cuidadosamente definidos, levando em consideração a inclusão das variáveis obrigatórias e a adoção de estratégias apropriadas para a realização dos experimentos.

Em seguida, os cenários foram estabelecidos para a implementação dos modelos, com o objetivo de obter resultados preditivos em relação à probabilidade de ocorrência de RDC

associado à SM. Essa estrutura de cenários permitiu a exploração de diferentes abordagens e configurações dos algoritmos, visando alcançar os melhores resultados de predição possíveis.

Dessa forma, a abordagem adotada proporcionou uma análise abrangente e detalhada, permitindo uma comparação eficiente entre os modelos e a identificação das estratégias mais eficazes para prever o RDC relacionado à SM.

6.3.2 Cenário 1

Inicialmente, procedeu-se à implementação de um algoritmo para avaliar a relevância dos marcadores da SM e os marcadores utilizados no FRS no que se refere à previsão do risco de desenvolvimento de DCV. O objetivo deste estudo foi identificar o RDC em uma população específica, realizando uma comparação entre indivíduos diagnosticados com SM e aqueles sem essa condição. Para tanto, utilizou-se o conjunto de dados "mswithinsulin2"proveniente do *National Health and Nutritional Evaluation Survey* (NHANES), composto por 29 colunas e 1.943 linhas. Esse conjunto de dados engloba informações de 693 pacientes diagnosticados com SM e 1.250 pacientes sem essa condição clínica. Dentre esses, 993 pacientes são do sexo feminino e 950 são do sexo masculino. Os marcadores empregados nesse conjunto de dados estão detalhadamente descritos na Tabela 1, a qual abrange variáveis antropométricas, bioquímicas e hemodinâmicas medidas no estudo. Esses marcadores foram utilizados para identificar a presença de SM nos participantes. Este experimento, assim como evidências presentes na literatura, corrobora a associação da SM com um risco elevado de DCV.

Durante a análise do conjunto de dados, constatou-se que a maioria dos marcadores presentes no FRS estava disponível, com exceção do colesterol total. Com base na tabela fornecida pelo pelo Ministério da Saúde, disponível em: [3], um algoritmo foi desenvolvido para atribuir pontos aos seis marcadores restantes e calcular o RDC dos indivíduos nesse conjunto de dados.

Após a estratificação dos dados, três colunas adicionais foram acrescentadas ao novo DataFrame, a saber: RiscoECV (Evento Cardiovascular Maior), Categoria e Doença, com o propósito de realizar análises estatísticas que contribuíssem para a compreensão do impacto da relação entre a SM e o RDC. Alguns estudos relacionados já mencionaram essa relação, como demonstrado na primeira diretriz brasileira, entretanto, até então, não havia um escore de RDC específico para essa população ou para essa condição diagnosticada nos pacientes. A Figura 3

[3] https://linhasdecuidado.saude.gov.br/portal/obesidade-no-adulto/unidade-de-atencao-primaria/planejamento-terapeutico/escore-risco-global-framingham/ acessado: em 13 de junho de 2023

representa o DataFrame resultante, que exibe essas novas colunas e os dados relacionados aos modelos de SM e RFS.

```
# Essa é a forma mais comum de pegar um subconjunto (das colunas) de um DataFrame:
# 6 Marcadores utilizados no Score de Framingham
df = sm[['age', 'sex', 'HDL', 'Hypertension', 'BloodGlucose', 'smoking', 'MetabolicSyndrome']]
df.head(2)
```

	age	sex	HDL	Hypertension	BloodGlucose	smoking	MetabolicSyndrome
0	22	Male	41	0	92	Never_smoker	0
1	44	Female	28	0	82	Never_smoker	0

[DataFrame com Marcadores de Risco de Framingham]

```
df.head(3)
```

	age	sex	HDL	Hypertension	BloodGlucose	smoking	MetabolicSyndrome	RiscoECV	Categoria	Doença
0	22	Male	41	0	92	Never_smoker	0	1.6	Baixo	Sem doença cardiovascular
1	44	Female	28	0	82	Never_smoker	0	1.2	Baixo	Sem doença cardiovascular
2	21	Male	43	0	107	Never_smoker	0	1.6	Baixo	Sem doença cardiovascular

[DataFrame com Três Colunas Novas]

Figura 3 – DataFrames dos Modelos de SM e RFS

No estudo de Framingham, que é considerado um dos estudos mais influentes na área da saúde cardiovascular, diversas doenças cardíacas, como a Doença Cardíaca Isquêmica, Insuficiência Cardíaca, Arritmia Cardíaca e Cardiomiopatia, estão associadas a diferentes níveis de risco. A escolha da doença específica a ser considerada depende do foco e dos objetivos do estudo em questão. Por exemplo, a Doença Cardíaca Isquêmica é relevante para analisar o risco moderado, enquanto a Hipertensão Arterial está associada a um risco mais baixo. A adoção de medidas preventivas, como a adoção de um estilo de vida saudável e o monitoramento regular, desempenha um papel crucial na redução do risco dessas condições (National Heart, Lung, and Blood Institute, 1948).

A abordagem utilizada neste estudo apresenta um potencial promissor para prever o RDC em indivíduos diagnosticados com SM. Os resultados obtidos neste cenário inicial serviram como base para o desenvolvimento de experimentos nos próximos cenários, nos quais técnicas de ML foram aplicadas para recomendar a melhor abordagem de análise e de predição do RDC associado à SM. O objetivo é contribuir para o desenvolvimento de um modelo ou escore de predição e diagnóstico da SM, bem como do RDC relacionado a ela.

Com base nos resultados obtidos na primeira etapa do estudo e após a realização das análises estatísticas, observa-se que a SM pode desempenhar um papel importante não apenas no desenvolvimento da própria síndrome, mas também no RDC. Nesse sentido, recomenda-se que o FRS, juntamente com variáveis adicionais, seja investigado em futuros estudos, visando validar, testar e aprimorar sua eficácia como ferramenta de avaliação. É essencial que essa investigação

seja conduzida por uma equipe multidisciplinar em colaboração com instituições hospitalares e laboratórios, a fim de obter resultados mais robustos e confiáveis.

6.3.3 Cenário 2

Após a condução dos testes necessários no primeiro cenário, constatou-se a viabilidade da aplicação de técnicas de aprendizado de máquina utilizando o mesmo conjunto de dados para predizer a SM e, por consequência, desenvolver a estimativa de RDC do indivíduo, tendo a SM como variável independente e os marcadores do FRS como variáveis dependentes presentes nesse conjunto de dados.

Considerando essas constatações e almejando atingir os objetivos estabelecidos neste estudo, foi desenvolvido um algoritmo que utiliza as técnicas mencionadas na subseção 6.2.1 "Ambiente de Implementação e Ferramentas Utilizadas"para selecionar o modelo de aprendizado de máquina mais apropriado para a predição da SM. Esse processo de seleção utiliza os marcadores de Framingham, os quais são apresentados na Tabela 4 de Marcadores de Framingham, como base para conduzir os testes e treinamentos dos modelos selecionados.

Tabela 4 – Marcadores de Framingham

Marcadores de Framingham
Sexo
Idade
HDL-C
Colesterol total
Pressão arterial sistólica (PAS) não tratada
Pressão arterial sistólica (PAS) tratada
Fumo
Diabetes
Escore de Risco Global (ERG) de Framingham

Fonte: E score de risco global (ERG) de Framingham: https://linhasdecuidado.saude.gov.br/portal/obesidade-no-adulto/unidade-de-atencao-primaria/planejamento-terapeutico/escore-risco-global-framingham/)

6.3.4 Cenário 3

No cenário 3, foram realizados os mesmos procedimentos adotados no cenário 2, porém utilizando apenas as variáveis relacionadas à SM. Essas variáveis são determinadas de acordo com o critério NCEP-ATP III, elaborado por um painel de especialistas em saúde dos Estados Unidos com o propósito de diagnosticar e tratar dislipidemias e outras condições relacionadas à saúde cardiovascular. É relevante ressaltar que a Organização Mundial da Saúde (OMS) e outras

instituições internacionais adotam critérios diagnósticos semelhantes para a SM, os quais incluem a presença de pelo menos três dos componentes listados na Tabela 5 de Critérios Diagnósticos para a SM, conforme estabelecido pelo NCEP-ATP III.

É fundamental destacar que as diretrizes clínicas para o diagnóstico e tratamento da SM estão sujeitas a atualizações ao longo do tempo, sendo essencial que os profissionais de saúde se mantenham atualizados com as últimas recomendações, a fim de oferecer a melhor assistência aos pacientes.

Tabela 5 – Critérios Diagnósticos para a SM segundo o (NCEP-ATP III)

Componentes	**Níveis**
Obesidade abdominal	Homens $>$ 102 cm, Mulher $>$ 88 cm
Triglicerídeos	$\geq$ 150 mg/dL
HDL Colesterol	Homens $\leq$ 40mg/dL, Mulher $\leq$ 50 mg/dL
Pressão arterial	Sistólica $\geq$ 130 mmHg, Diastólica $\geq$ 85 mmHg
Glicemia de jejum	$\geq$ 110 mg/dL

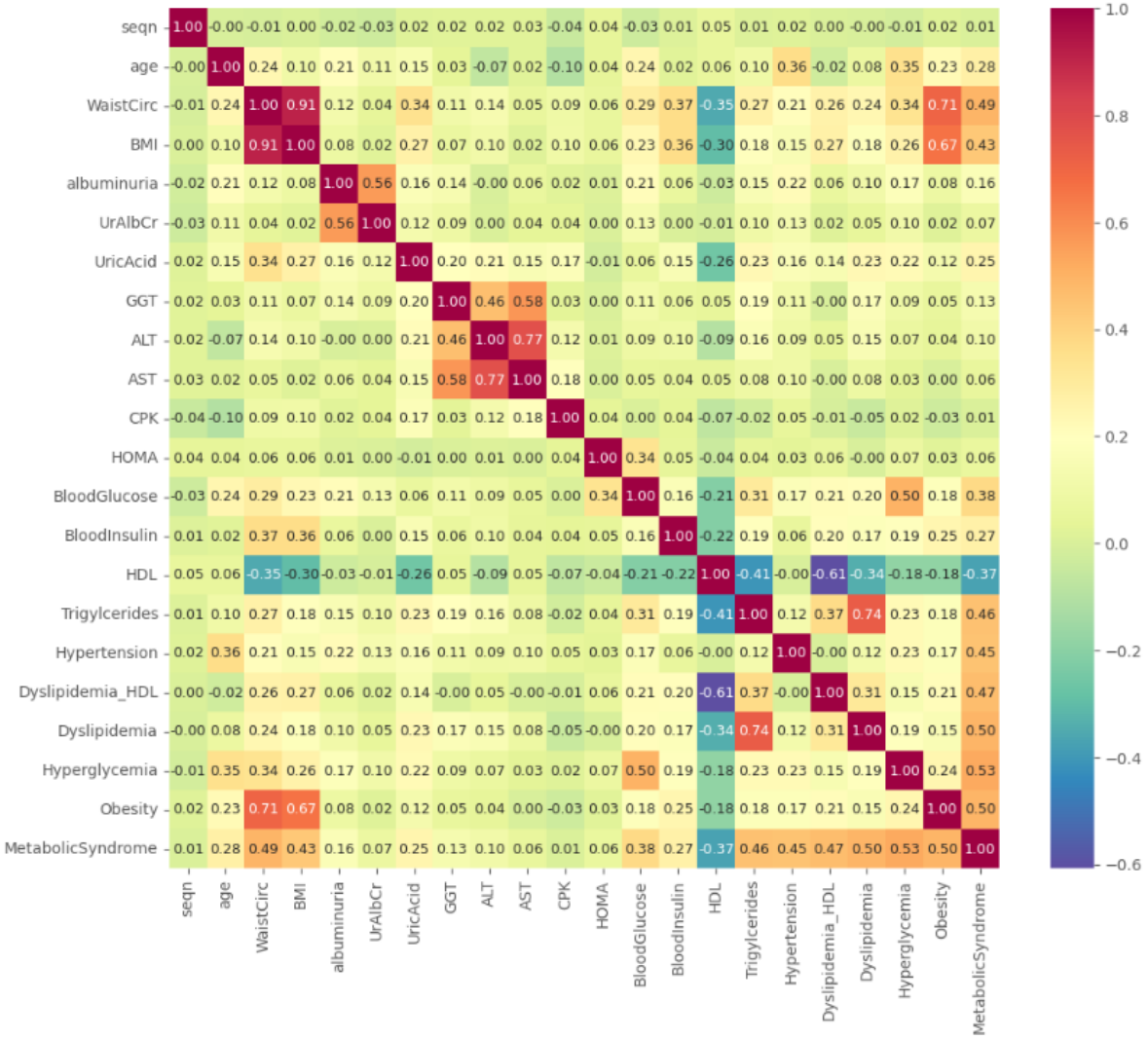
Fonte: Arquivos Brasileiros de Cardiologia

Fonte: : Elaboração própria (2023)

6.3.5 Cenário 4

Após uma análise individual dos resultados obtidos nos cenários 2 e 3, foi conduzido um estudo para investigar as relações entre as variáveis por meio de um mapa de calor. O objetivo principal dessa etapa foi identificar relações interessantes que evidenciassem a influência de uma variável sobre a outra, especialmente em relação ao alvo em estudo. Para alcançar esse objetivo, utilizou-se a biblioteca Seaborn para criar uma matriz de correlação que permitiu identificar as variáveis mais correlacionadas. Essa técnica foi aplicada durante o processo de normalização dos dados, como ilustrado na Figura 4.

Figura 4 – Matriz de Correlação



Fonte: : Elaboração própria (2023)

As relações descobertas revelam informações relevantes sobre a influência mútua das variáveis, especialmente em relação ao alvo em estudo. A matriz de calor apresenta as correlações entre as variáveis, permitindo uma visualização clara da intensidade dessas relações. A interpretação das relações leva em consideração a magnitude dos coeficientes de correlação, em que valores próximos de 1 indicam uma correlação positiva forte, valores próximos de -1 indicam uma correlação negativa forte e valores próximos de 0 indicam uma correlação fraca ou inexistente (PEDREGOSA, 2011).

Essas descobertas podem fornecer *insights* valiosos para compreender a interação entre as variáveis relacionadas à SM e ao FRS, auxiliando na identificação de RDC e no desenvolvimento de modelos preditivos mais precisos. Além da análise das relações por meio da matriz de correlação, foi implementado um código para treinar um modelo de regressão usando o algoritmo

de Floresta Aleatória, selecionado como o melhor modelo para este experimento (PEDREGOSA, 2011). O objetivo desse modelo é prever a variável dependente "MetabolicSyndrome"com base nas variáveis independentes presentes no DataFrame "df".

Com base nessa análise exploratória, foram realizados experimentos que consistiram na integração das variáveis utilizadas no diagnóstico da SM com as variáveis do FRS, com o objetivo de selecionar o modelo mais adequado para a predição. Dentro desse contexto, foram executados procedimentos semelhantes, mas direcionados a modelos de classificação, a fim de determinar se um indivíduo apresenta ou não a condição clínica esperada. Os resultados desses experimentos serão detalhados na seção de resultados, enquanto o código implementado será disponibilizado nos apêndices correspondentes a cada cenário específico.

6.3.6 Cenário 5

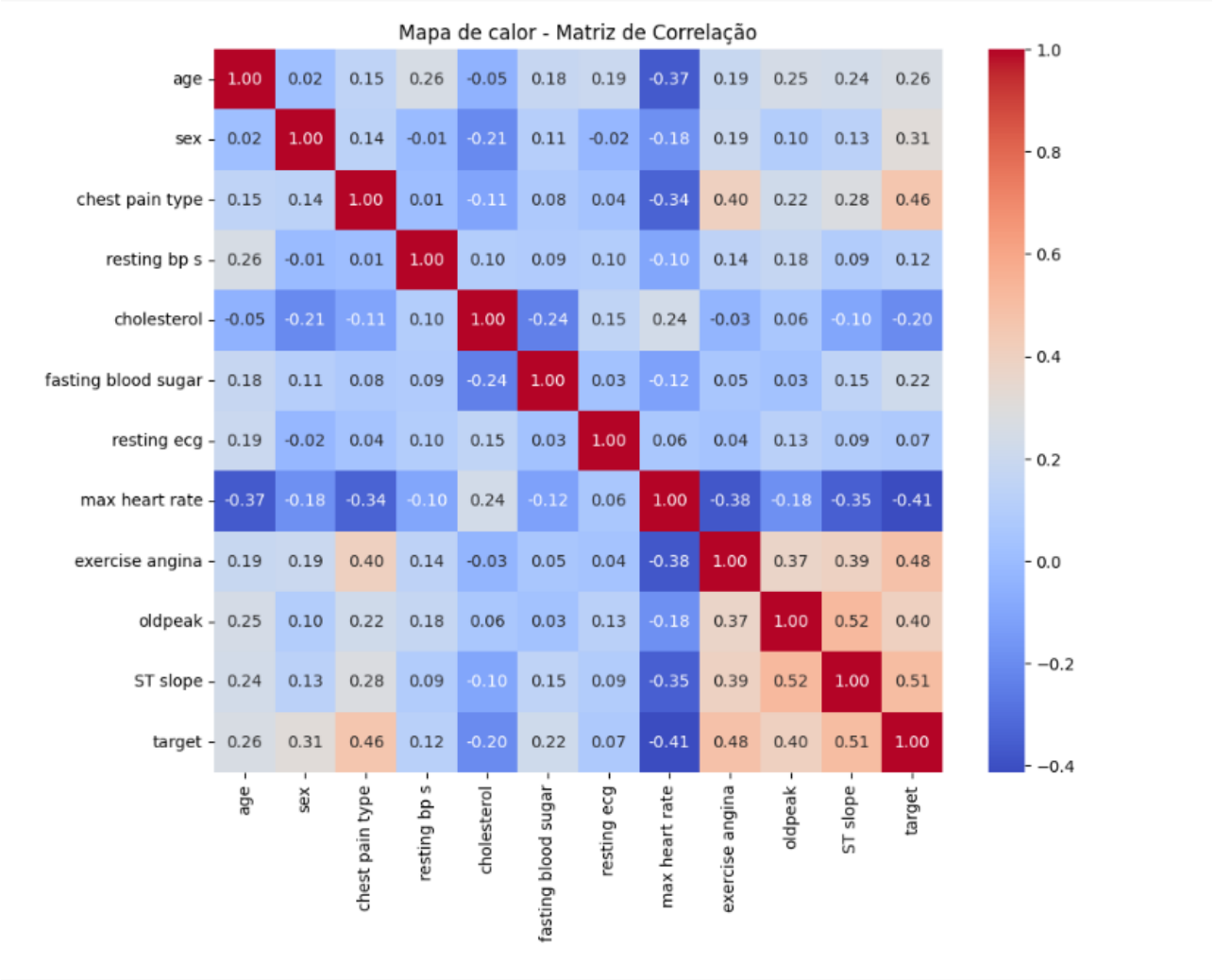
Nesse contexto, foi utilizado os mesmos procedimentos realizados nos cenários 2, 3 e 4, empregando um conjunto de dados combinados de doenças cardíacas provenientes de pacientes de Cleveland-EUA, Hungria, Suíça e VA Long Beach. Os Marcadores Clínicos desse conjunto de dados estão disponíveis na Tabela 2.

A população do conjunto de dados consiste em um total de 1190 indivíduos, sendo 909 do sexo masculino e 281 do sexo feminino. Seguiremos os passos mencionados na seção 6.2, "Abordagem para o Desenvolvimento e Seleção do Modelo Recomendado", para selecionar o melhor modelo nesse conjunto de dados.

A fim de escolher o modelo de aprendizado mais adequado e suas respectivas combinações de variáveis da SM e Framingham para prever o RDC, será necessário ajustar um modelo aos dados e avaliar seu desempenho utilizando métricas apropriadas.

Dessa forma, poderemos identificar o modelo mais adequado para prever o RDC com base nas variáveis selecionadas da SM e FRS, seguindo um processo estruturado e robusto de análise e seleção de modelos. Para alcançar esse objetivo, iniciaremos analisando a importação e visualização dos dados. Utilizamos a biblioteca Seaborn para criar uma matriz de correlação, permitindo identificar as variáveis mais correlacionadas. Essa técnica foi aplicada durante o processo de normalização dos dados, como ilustrado na Figura 5.

Figura 5 – Matriz de Correlação DataSet: Cleveland, Hungria, Suíça e VA Long Beach



Fonte: : Elaboração própria (2023)

A matriz de correlação é uma representação tabular que mostra as correlações entre pares de variáveis em um conjunto de dados. Cada célula da matriz contém um valor que varia de -1 a 1, indicando a força e a direção da correlação.

Valores próximos de 1 indicam uma correlação positiva forte, o que significa que as variáveis tendem a se mover na mesma direção. Por outro lado, valores próximos de -1 indicam uma correlação negativa forte, indicando que as variáveis tendem a se mover em direções opostas. Valores próximos de 0 indicam uma correlação fraca ou inexistente entre as variáveis.

Ao analisar o mapa de calor da matriz de correlação, é possível identificar padrões e relações entre as variáveis. Variáveis fortemente correlacionadas podem indicar uma dependência mútua ou uma relação causal entre elas. Por exemplo, se houver uma alta correlação positiva entre a pressão arterial e o colesterol, isso sugere que essas variáveis estão relacionadas e podem afetar uma à outra.

Com base na matriz de correlação fornecida, podemos resumir a relação das variáveis da SM e Framingham em relação ao alvo (target) das seguintes formas:

- **Variáveis com correlação fraca com o alvo:**
  - Idade (age): A idade apresenta uma correlação fraca positiva com o alvo. Isso significa que, em média, o aumento da idade está levemente associado a um aumento no RDC.
- **Variáveis com correlação moderada com o alvo:**
  - Sexo (sex): O sexo apresenta uma correlação moderada positiva com o alvo. Isso indica que existe uma relação um pouco mais significativa entre o sexo e o RDC, sugerindo que os homens podem ter um risco ligeiramente maior do que as mulheres.
  - Tipo de dor no peito (chest pain type): O tipo de dor no peito apresenta uma correlação moderada positiva com o alvo. Isso sugere que certos tipos de dor no peito podem estar associados a um aumento do RDC.
  - Angina induzida por exercício (exercise angina): A presença de angina induzida por exercício apresenta uma correlação moderada positiva com o alvo. Isso indica que a presença dessa condição está relacionada a um maior RDC.
- **Variáveis com correlação forte com o alvo:**
  - Oldpeak: A depressão do segmento ST induzida pelo exercício apresenta uma correlação forte positiva com o alvo. Isso significa que um maior valor de depressão do segmento ST está fortemente associado a um aumento do RDC.

É importante ressaltar que a força da correlação pode variar dependendo do contexto e dos critérios de interpretação. A análise realizada aqui é baseada exclusivamente na magnitude das correlações presentes na matriz fornecida.

Os resultados desses experimentos serão detalhados na seção de resultados, enquanto o código implementado será disponibilizado nos apêndices correspondentes a cada cenário específico.

6.3.7 Cenário 6

O terceiro conjunto de dados utilizado neste projeto é o Framingham, que é composto por mais de 4.240 registros, sendo 2.420 do sexo feminino e 1.820 do sexo masculino. Entre esses registros, 2.095 correspondem a fumantes e 2.145 são não fumantes. Esse conjunto

de dados possui 16 colunas e fornece informações relevantes sobre 15 atributos. O arquivo "framingham.csv"contém os dados coletados no renomado Estudo de Framingham, uma pesquisa epidemiológica sobre DCV realizada em Framingham, Estados Unidos, a partir de 1948. O objetivo desse estudo foi identificar os principais fatores de risco associados a doenças cardíacas (Ministério da Saúde, 13 de junho de 2023).

Durante o processo de normalização dos dados, foi observado que o conjunto de dados continha um número considerável de valores ausentes. Por esse motivo, optou-se por remover as linhas que continham esses valores ausentes, visando garantir a integridade dos dados utilizados para fins acadêmicos e de pesquisa. Após essa remoção, o conjunto de dados foi reduzido para 3658 observações (linhas) e 16 variáveis (colunas), indicando que 582 linhas foram excluídas devido à presença de pelo menos um valor ausente.

É importante ressaltar que a estratégia de exclusão das linhas com valores ausentes pode ser adequada, dependendo do contexto e dos objetivos da análise. No entanto, é necessário considerar que essa exclusão pode acarretar na perda de informações e possivelmente introduzir viés nos resultados da análise. Portanto, é fundamental avaliar cuidadosamente os efeitos dessa remoção nos dados e nas análises subsequentes.

Antes da exclusão das linhas com valores ausentes, o conjunto de dados continha 2420 registros para indivíduos do sexo feminino (0) e 1820 registros para indivíduos do sexo masculino (1), assim como 2145 registros para não fumantes (0) e 2095 registros para fumantes (1). Após a exclusão, verificou-se que o conjunto de dados reduzido contém 2035 registros para indivíduos do sexo feminino e 1623 registros para indivíduos do sexo masculino, assim como 1869 registros para não fumantes e 1789 registros para fumantes. Essa diferença na contagem de registros reflete o impacto da exclusão das linhas com valores ausentes nos dados, incluindo a distribuição dos registros entre os sexos. Seguindo a mesma abordagem de apresentação dos resultados, o código do modelo está disponível no respectivo apêndice do cenário 6, e os resultados são apresentados na seção de resultados.

7 ANÁLISE DOS RESULTADOS DOS MODELOS DE REGRESSÃO E CLASSIFICAÇÃO NA PREDIÇÃO DE RDC ASSOCIADO A SM

O objetivo desta pesquisa é desenvolver e avaliar modelos de regressão e classificação capazes de prever o RDC associado à presença da SM. A SM, caracterizada pela coexistência de fatores de RDCes como obesidade central, hipertensão arterial, dislipidemia e resistência à insulina, tem sido reconhecida como um importante preditor de complicações cardiovasculares em indivíduos. Nesse sentido, identificar e quantificar adequadamente o RDC relacionado à SM é fundamental para uma intervenção precoce e eficaz.

No entanto, apesar da relevância clínica dessa associação, ainda há lacunas no conhecimento sobre o desenvolvimento de escore e/ou modelos de aprendizado de máquina específicos para a predição do RDC em pacientes com SM (BRANDÃO, 2005) e (PRECOMA, 2019). Dessa forma, esta pesquisa busca preencher essa lacuna, desenvolvendo modelos robustos e avaliando sua capacidade de prever com precisão o RDC em pacientes com SM.

Ao estabelecer esses modelos preditivos, espera-se que este estudo forneça *insights* valiosos para a identificação precoce de indivíduos com maior RDC, permitindo intervenções preventivas personalizadas e direcionadas. Além disso, os resultados obtidos podem contribuir para aprimorar a estratificação de risco e a tomada de decisão clínica, auxiliando na redução da morbidade e mortalidade associadas às complicações cardiovasculares na população afetada pela SM.

Para embasar essa pesquisa, serão utilizadas informações de estudos prévios sobre a prevalência da SM e seus fatores de RDCes, bem como referências a modelos de regressão e predição em outros contextos relacionados à saúde cardiovascular. Essas referências fornecerão um alicerce sólido para a pesquisa e demonstrarão a relevância e a contribuição original deste estudo para o campo da medicina cardiovascular e metabólica.

Nesta seção, serão apresentados os resultados obtidos em relação aos cenários propostos, bem como a comparação com os modelos de aprendizagem de máquina implementados nesta pesquisa. É importante destacar que, devido à normalização dos dados, houve um impacto significativo na quantidade de pacientes do banco de dados do estudo de Framingham. Foram identificadas duplicações, valores ausentes e/ou valores fora dos padrões estabelecidos para determinadas propostas do conjunto de dados.

Apesar dos desafios enfrentados, os resultados obtidos foram considerados satisfatórios. Os algoritmos propostos puderam ser aplicados com sucesso e sua descrição detalhada será apresentada a seguir, estando os códigos disponíveis no apêndice deste trabalho. Os resultados

serão apresentados de acordo com a sequência estabelecida para cada cenário de experimento realizado.

7.1 CENÁRIOS: RESULTADOS DOS EXPERIMENTOS

Com os dados devidamente normalizados e disponíveis para a implementação dos algoritmos, foram delineados cenários que atendessem às hipóteses e propostas iniciais, considerando a utilização das variáveis relevantes para o objetivo deste trabalho.

Nesse sentido, a estrutura dos cenários adotados compreende o Cenário 1, no qual foi aplicado o algoritmo de estratificação baseado na atribuição de pontos, seguindo como orientação o sistema de risco global disponibilizado pelo Ministério da Saúde. O objetivo desse cenário foi obter resultados preditivos em relação à probabilidade de um paciente apresentar RDC nos próximos 10 anos, conforme as pontuações atribuídas aos marcadores utilizados no FRS.

Nos Cenários 2, 3 e 4, foi empregado o mesmo conjunto de dados utilizado no Cenário 1, devido à inclusão das variáveis relevantes tanto para a SM quanto para o FRS. Nestes experimentos, foram aplicadas técnicas de algoritmos de aprendizado de máquina previamente mencionados na seção de metodologia experimental, com o propósito de estimar o RDC relacionado à SM.

Os Cenários 5 e 6 seguiram os mesmos procedimentos realizados nos Cenários 2, 3 e 4, respectivamente, porém utilizando conjuntos de dados distintos. Os resultados obtidos nesses cenários serão devidamente apresentados e explicados em suas respectivas seções.

7.2 ANÁLISE E RESULTADOS DO CENÁRIO 1: ESTATÍSTICAS DESCRITIVAS DO RDC NA POPULAÇÃO GERAL E NAS SUBPOPULAÇÕES COM E SEM SM

Neste estudo, adotou-se o algoritmo de estratificação baseado na atribuição de pontos, conforme as diretrizes estabelecidas pelo Ministério da Saúde, com o propósito de avaliar de forma precisa o RDC. Entretanto, o foco principal desta pesquisa foi a análise da associação entre a SM e o RDC. Para essa investigação, utilizou-se o conjunto de dados denominado "mswithinsulin2", proveniente do *National Health and Nutritional Evaluation Survey* (NHANES), que é composto por 29 colunas e 1.943 linhas. É importante ressaltar que esse conjunto de dados foi apresentado detalhadamente na seção de metodologia, especificamente no cenário 1 do experimento.

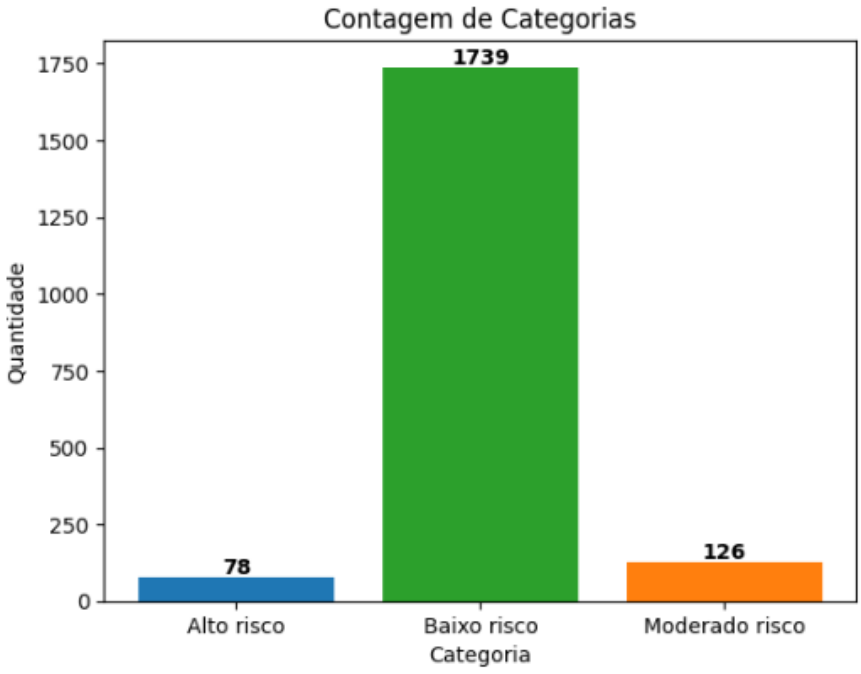
Embora este experimento em particular não tenha explorado o uso de técnicas de IA, a análise dos marcadores fornecerá informações valiosas sobre a relação entre à SM e o RDC.

A compreensão dessa relação é fundamental para o avanço do conhecimento na área e para o desenvolvimento de estratégias de prevenção e tratamento mais eficazes.

7.2.1 Apresentação e Análise dos Resultados Obtidos

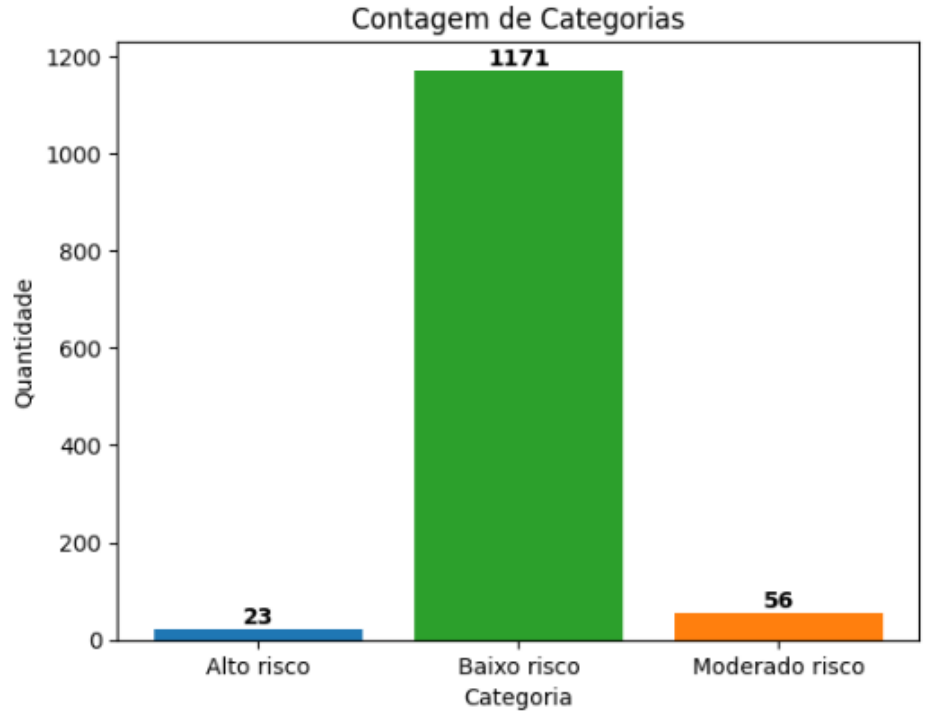
Apesar da não recomendação de utilização do FRS para prever o RDC nessa população, conforme destacado por (BRANDÃO, 2005), foi possível identificar os pacientes com risco baixo, moderado ou alto por meio dos gráficos, mesmo considerando a ausência do colesterol total como indicador no escore de FRS desse conjunto de dados. A FIGURA 6 apresenta o resultado do algoritmo aplicado à população geral do conjunto de dados, evidenciando a distribuição dos pacientes em relação aos diferentes níveis de risco. Além disso, a FIGURA 7 exibe o resultado do algoritmo aplicado à população sem SM do conjunto de dados, permitindo a visualização da distribuição do RDC nesse subgrupo específico. Por fim, a FIGURA 8 ilustra o resultado do algoritmo aplicado à população com SM do conjunto de dados, proporcionando *insights* sobre a associação entre essa condição clínica e o RDC. Essas informações são fundamentais para compreender a relação entre os fatores de risco e o desenvolvimento de eventos cardiovasculares nessa população específica.

Figura 6 – Quantidade por categoria da população geral do Dataset com risco Alto, Moderado e Baixo



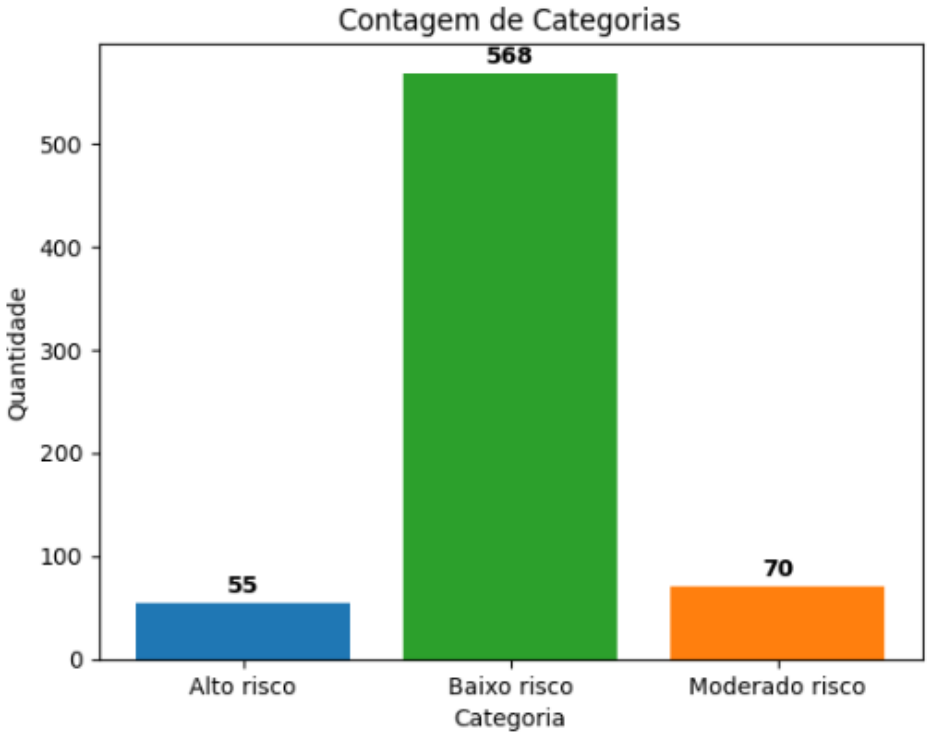
Fonte: Elaboração própria (2023)

Figura 7 – Quantidade por categoria da população sem SM do Dataset com risco Alto, Moderado e Baixo



Fonte: Elaboração própria (2023)

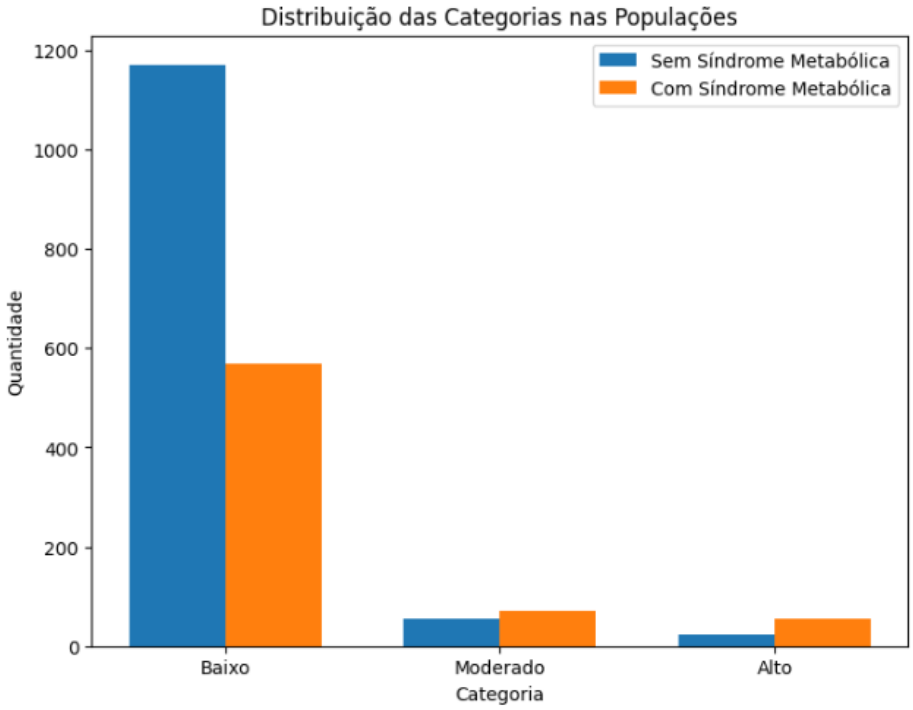
Figura 8 – Quantidade por categoria da população com SM do Dataset com risco Alto, Moderado e Baixo



Fonte: Elaboração própria (2023)

A FIGURA 9 ilustra a distribuição das categorias de RDC nas populações analisadas, oferecendo *insights* valiosos sobre a proporção de indivíduos em cada categoria.

Figura 9 – Distribuição das Categorias nas Populações



7.2.2 Resultados e Avaliação do Cenário 1: Estatísticas Descritivas do RDC nas Populações Geral, sem SM e com SM

As estatísticas descritivas do RiscoECV nas diferentes populações revelaram informações relevantes sobre a distribuição desse RDC. Na população geral, composta por 1943 indivíduos, a média do RiscoECV foi de 3.974, com um desvio padrão de 6.018. A menor pontuação registrada foi de 0.5, enquanto a maior foi de 30. As estatísticas de quartis mostraram que 25% dos indivíduos apresentaram um RDC abaixo de 1.2, 50% tiveram um risco abaixo de 1.6 e 75% apresentaram um risco inferior a 3.3.

Ao analisar a população sem SM, composta por 1250 indivíduos, constatou-se que a média do RiscoECV foi de 3.077, com um desvio padrão de 4.469. As estatísticas de quartis para essa população foram semelhantes às da população geral. Já na população com SM, composta por 693 indivíduos, a média do RiscoECV foi de 5.592, com um desvio padrão de 7.844. A pontuação mínima registrada foi de 1, enquanto a máxima foi de 30. Os quartis revelaram que 25% dos indivíduos tiveram um RDC abaixo de 1.2, 50% abaixo de 1.6 e 75% abaixo de 5.3.

Ao comparar as médias do RiscoECV entre as populações, observou-se uma diferença percentual de 81.74%. Essa diferença indica que a população com SM apresentou uma média de

RDC significativamente maior em comparação com a população geral.

No grupo, que representa a população masculina, foram calculadas as estatísticas descritivas para a variável "RiscoECV", que mede o RDC. Com base nos resultados, a média de RDC foi de 5.851, com um desvio padrão de 7.709. A menor medida de risco foi de 1.100 e a maior medida foi de 30.000. A distribuição dos dados revelou que 25% dos indivíduos apresentaram um RDC abaixo de 1.600, enquanto 50% apresentaram um risco abaixo de 1.900 e 75% abaixo de 5.600.

Já no grupo, que representa a população feminina, também foram calculadas as estatísticas descritivas para a variável "RiscoECV". A média de RDC encontrada foi de 2.178, com um desvio padrão de 2.732. A menor medida de risco foi de 0.500 e a maior medida foi de 24.800. A distribuição dos dados revelou que 25% das mulheres apresentaram um RDC abaixo de 1.200, enquanto 50% apresentaram um risco abaixo de 1.200 e 75% abaixo de 2.000.

Esses resultados indicam que, na amostra analisada, os homens apresentaram, em média, um RDC maior em comparação com as mulheres. Além disso, a dispersão dos valores de risco também foi maior no grupo masculino, evidenciada pelo desvio padrão mais elevado.

É importante ressaltar que esses resultados são específicos para a amostra analisada e não podem ser generalizados para toda a população. No entanto, eles fornecem indícios de que o gênero pode desempenhar um papel na predisposição ao RDC na SM. Estudos adicionais são necessários para uma compreensão mais abrangente dessa relação e para identificar os fatores específicos que contribuem para essas diferenças de risco entre homens e mulheres.

Esses resultados sugerem que a presença da SM está associada a um aumento do RDC. Essa informação é relevante para o desenvolvimento de estratégias de prevenção e tratamento direcionadas a essa população específica, visando reduzir os potenciais problemas cardiovasculares associados à SM.

7.3 ANÁLISE E RESULTADOS DO CENÁRIO 2: SELEÇÃO DO MELHOR MODELO DE ML UTILIZANDO OS MARCADORES DO FRS

7.3.1 Introdução e Contextualização do Cenário de Resultados Obtidos

Em conformidade com os procedimentos recomendados na seção de metodologia do experimento, buscamos desenvolver um modelo de aprendizado de máquina para selecionar o melhor modelo de regressão na predição da SM. Como mencionado anteriormente, a SM é considerada um preditor para determinar se um paciente possui um risco elevado de desenvolvê-la e, consequentemente, um RDC associado a essa condição. Além disso, também foi desenvolvido

um modelo de classificação com as mesmas técnicas para o diagnóstico da presença da SM.

No contexto da regressão, o modelo RF foi identificado como o mais adequado neste cenário. Utilizando apenas as variáveis disponíveis no FRS, esse modelo obteve o melhor desempenho na predição da SM. Vale ressaltar que, embora a variável de colesterol total não estivesse presente no conjunto de dados, as outras seis variáveis utilizadas no FRS desempenharam um papel importante nos experimentos e permitiram atingir o objetivo desta pesquisa. A análise detalhada dos resultados, considerando métricas de desempenho específicas, proporcionará uma interpretação mais precisa e conclusões relevantes acerca do modelo de regressão.

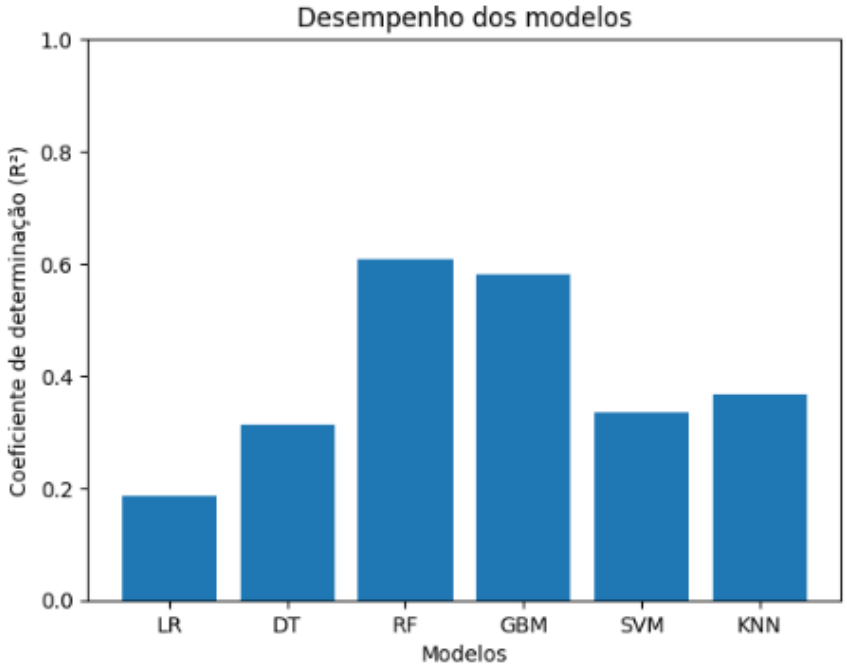
Por outro lado, no que diz respeito à classificação, o modelo RandomForestClassifier, também utilizando as mesmas variáveis do FRS, foi selecionado como o mais adequado para diagnosticar a presença da SM. A avaliação do desempenho desse modelo, considerando métricas específicas de classificação, fornecerá informações valiosas sobre a capacidade de identificação correta da presença ou ausência da SM.

Nos próximos trechos, apresentaremos os resultados obtidos para o modelo de regressão, fornecendo detalhes sobre as métricas de desempenho e interpretação dos resultados. Em seguida, abordaremos os resultados encontrados para o modelo de classificação, com uma análise similar. Essa abordagem nos permitirá tirar conclusões mais robustas e relevantes acerca da capacidade preditiva e diagnóstica dos modelos desenvolvidos neste estudo.

7.3.1.1 Seleção do Melhor Modelo de Regressão

O primeiro passo consistiu na criação de pipelines para os modelos utilizados nesta pesquisa, com o objetivo de selecionar o melhor entre eles. Os modelos considerados foram os seguintes: LR, DT, RF, GBM, SVM e KNN. Cada modelo foi incorporado em um pipeline, sendo identificado como 'Regressor'. Através da avaliação comparativa, buscou-se identificar o modelo com melhor desempenho para a predição da SM. A FIGURA 10 apresenta o gráfico que detalha o melhor modelo obtido.

Figura 10 – Desempenho dos Modelos de Regressão Utilizando Marcadores de Framingham



Fonte: Elaboração própria (2023)

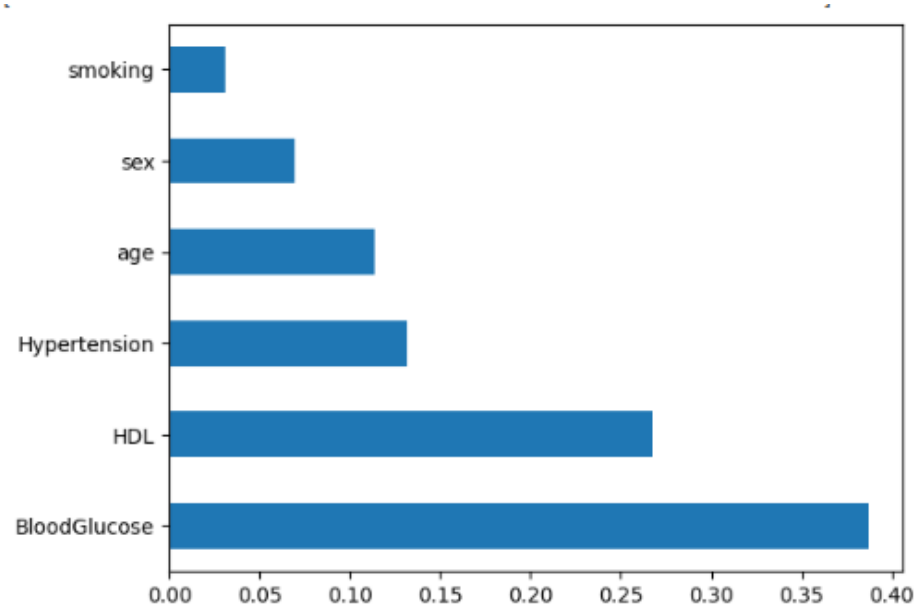
Após a seleção do melhor modelo, a etapa de treinamento é realizada utilizando o modelo escolhido, neste caso o RandomForestRegressor(). Para isso, uma instância do modelo é criada e as variáveis independentes (X) e a variável dependente (y) são separadas do DataFrame "data". Em seguida, o modelo é ajustado aos dados de treinamento por meio da função fit().

Uma vez concluído o treinamento, são avaliadas as importâncias das características (feature-importances) fornecidas pelo modelo. Essas importâncias revelam o grau de contribuição de cada variável independente na previsão da variável dependente. É importante destacar que essas importâncias são calculadas com base na análise do modelo e refletem a relevância das características no processo de predição.

As importâncias das características são então impressas, permitindo uma visualização clara e objetiva da contribuição de cada variável independente. Esse aspecto é fundamental para compreender quais características têm maior influência na previsão da variável dependente e auxiliar na interpretação dos resultados obtidos pelo modelo.

Dessa forma, a análise das importâncias das características proporciona *insights* valiosos sobre quais variáveis têm maior poder de explicação e influência no processo de previsão, contribuindo para o entendimento do funcionamento do modelo e para a tomada de decisões embasadas nos resultados obtidos. O gráfico resultante é exibido utilizando a função plt.show(), como demonstrado na FIGURA11.

Figura 11 – Gráfico de Importância das Características de Regressão Utilizando Marcadores de Framingha



Fonte: Elaboração própria (2023)

O modelo escolhido é o RF, e os resultados são expostos por meio de métricas de desempenho, que englobam o R2 Score, Mean Absolute Error, Mean Squared Error e Root Mean Squared Error. A tabela 6 exibe os resultados alcançados pelo modelo mais bem-sucedido neste experimento.

O R2 Score é uma medida de quão bem o modelo se ajusta aos dados, variando de 0 a 1, sendo 1 o ajuste perfeito. No caso do modelo RF, o R2 Score obtido foi de 0.627778, indicando que aproximadamente 62,78% da variabilidade dos dados foi explicada pelo modelo.

O MAE é a média dos erros absolutos entre as previsões do modelo e os valores reais. Neste caso, o MAE foi de 0.166823, o que significa que, em média, as previsões do modelo apresentaram um desvio absoluto de aproximadamente 0.166823 unidades em relação aos valores reais.

O MSE é a média dos erros quadrados entre as previsões do modelo e os valores reais. Para o modelo RF, o MSE foi de 0.0834469, indicando que, em média, o quadrado das diferenças entre as previsões e os valores reais foi de aproximadamente 0.0834469 unidades.

O RMSE é a raiz quadrada do MSE e fornece uma medida do desvio padrão dos erros do modelo. No caso do modelo RF, o RMSE foi de 0.288872, o que significa que, em média, os erros do modelo apresentaram um desvio padrão de aproximadamente 0.288872 unidades em relação aos valores reais.

Esses resultados fornecem uma avaliação do desempenho do modelo RF na predição

da SM. Quanto mais próximo de 1 for o R2 Score e menor forem o MAE, o MSE e o RMSE, melhor será o desempenho do modelo. Portanto, com base nos resultados apresentados, o modelo RF mostrou-se promissor na tarefa de predição da SM neste experimento.

Tabela 6 – Resultados dos modelos de regressão

Modelo	**R2 Score**	**Mean Absolute Error**	**Mean Squared Error**	**RMSE**
RF	0.62	0.16	0.08	0.28

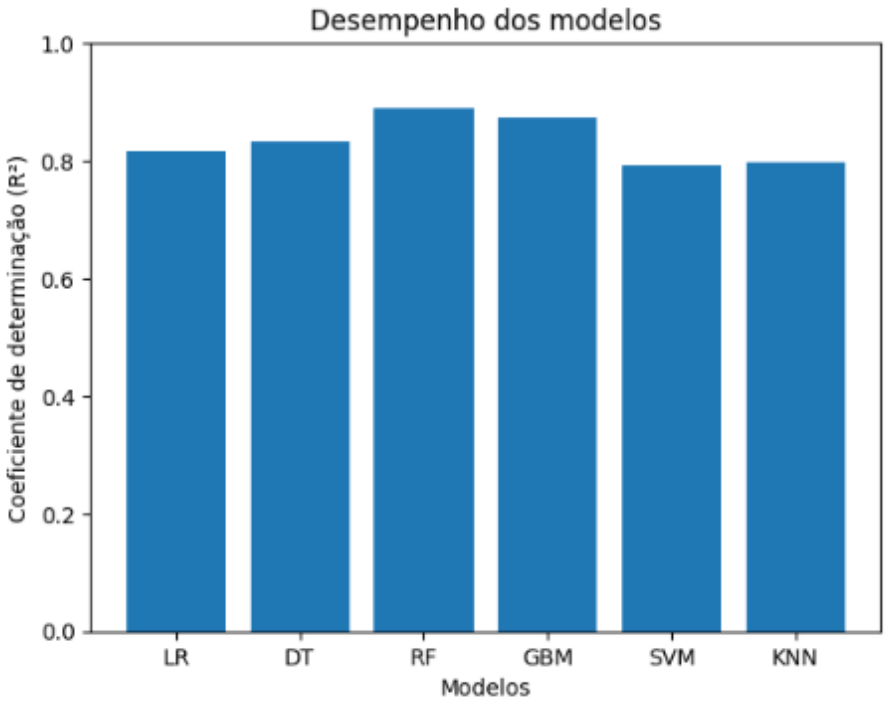
7.3.1.2 Seleção do Melhor Modelo de Classificação

Com o propósito de selecionar o modelo de classificação mais adequado para o diagnóstico da SM, os experimentos realizados nos quatro primeiros cenários incorporaram pipelines para cada um dos modelos considerados: LR, DT, RF, GBM, SVM e KNN.

Esses pipelines foram cuidadosamente projetados para garantir a padronização do pré-processamento dos dados e a aplicação consistente das etapas de treinamento e teste em cada modelo. Posteriormente, os modelos foram submetidos a uma avaliação abrangente utilizando métricas apropriadas para problemas de classificação, com o objetivo de identificar aquele que apresenta a melhor capacidade de diagnóstico da SM.

A implementação dos modelos de classificação é minuciosamente descrita no código fornecido no Apêndice A, na seção dedicada ao Cenário 2, que concentra-se na seleção do modelo de classificação mais eficaz para o diagnóstico da SM. A análise dos resultados obtidos será apresentada a seguir, fornecendo uma visão aprofundada do desempenho e das capacidades de diagnóstico de cada modelo. A FIGURA 12 apresenta o gráfico que detalha o melhor modelo de classificação obtido.

Figura 12 – Desempenho dos Modelos de Classificação Utilizando Marcadores de Framingham



Fonte: Elaboração própria (2023)

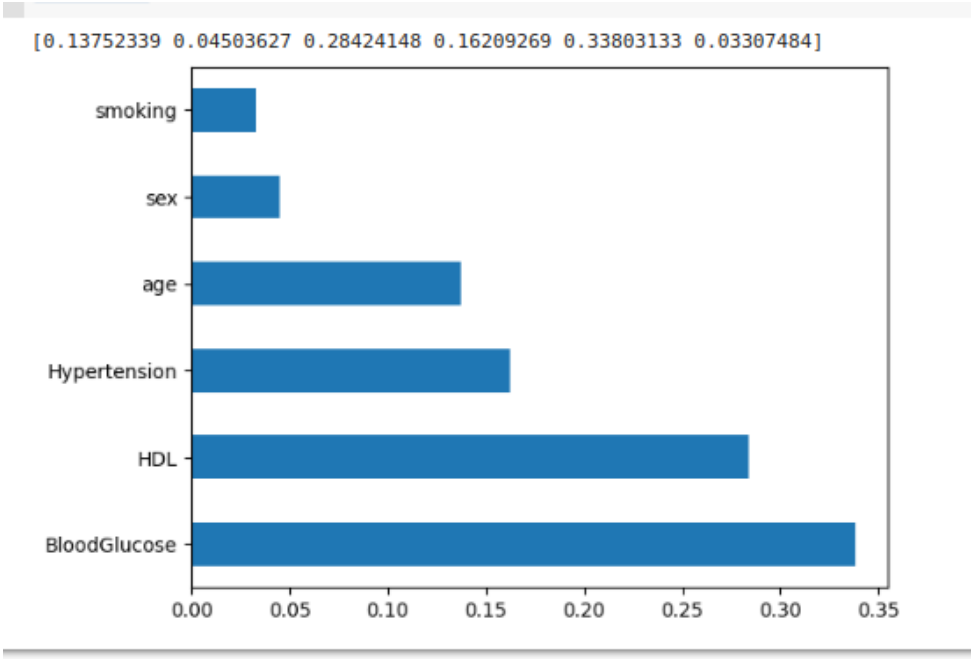
A fim de visualizar as importâncias das características nos modelos de classificação, o mesmo procedimento mencionado na seleção do melhor modelo de regressão foi aplicado. Utilizou-se o código que faz uso da biblioteca pandas para criar uma série denominada "feat-importances", a qual associa as importâncias das características aos nomes das colunas em X, representando as variáveis independentes.

Em seguida, foram selecionadas as características com as maiores importâncias e um gráfico de barras horizontais foi gerado para representar essas importâncias. Esse gráfico fornece uma representação visual das principais variáveis que contribuem para a previsão do diagnóstico da SM.

Para apresentar o resultado, utilizou-se a função plt.show(), possibilitando a exibição do gráfico resultante. Esse gráfico permite uma análise intuitiva das importâncias das características, facilitando a identificação das variáveis mais relevantes para o modelo de classificação.

A FIGURA 13 ilustra o gráfico gerado, demonstrando de forma clara e concisa as importâncias das características nos modelos de classificação. Essa análise auxilia na compreensão das variáveis que exercem maior influência na classificação da SM, contribuindo para a interpretação dos resultados obtidos e fornecendo *insights* relevantes para futuras investigações.

Figura 13 – Gráfico de Importância das Características de Classificação Utilizando Marcadores de Framingha



Fonte: Elaboração própria (2023)

Os resultados obtidos revelaram que o modelo de classificação RF demonstrou o melhor desempenho na tarefa de diagnóstico da SM. Os parâmetros ótimos encontrados para este modelo foram um valor de 'max-depth' igual a 7 e um número de estimadores ('n-estimators') igual a 100.

A acurácia alcançada pelo melhor modelo foi de 0.884, o que indica que ele foi capaz de classificar corretamente aproximadamente 88.4% das amostras. Ao analisar a métrica de precisão, verificou-se que o modelo obteve uma taxa de 0.90 para a classe 0 (indicando a ausência da SM) e 0.84 para a classe 1 (indicando a presença da SM). Esses valores demonstram que o modelo apresentou uma boa capacidade de identificar corretamente os casos negativos e positivos.

Ao avaliar a métrica de recall, que indica a taxa de verdadeiros positivos, observou-se um valor de 0.92 para a classe 0 e 0.81 para a classe 1. Isso indica que o modelo foi capaz de recuperar a grande maioria dos casos da classe 0, porém apresentou um desempenho ligeiramente inferior na identificação dos casos da classe 1.

A métrica de f1-score, que combina as métricas de precisão e recall, alcançou um valor de 0.91 para a classe 0 e 0.83 para a classe 1. Esses resultados indicam um bom equilíbrio entre precisão e recall para ambas as classes.

Em resumo, o modelo de classificação RF, utilizando os parâmetros 'max-depth' igual a 7 e 'n-estimators' igual a 100, demonstrou um desempenho satisfatório no diagnóstico da SM.

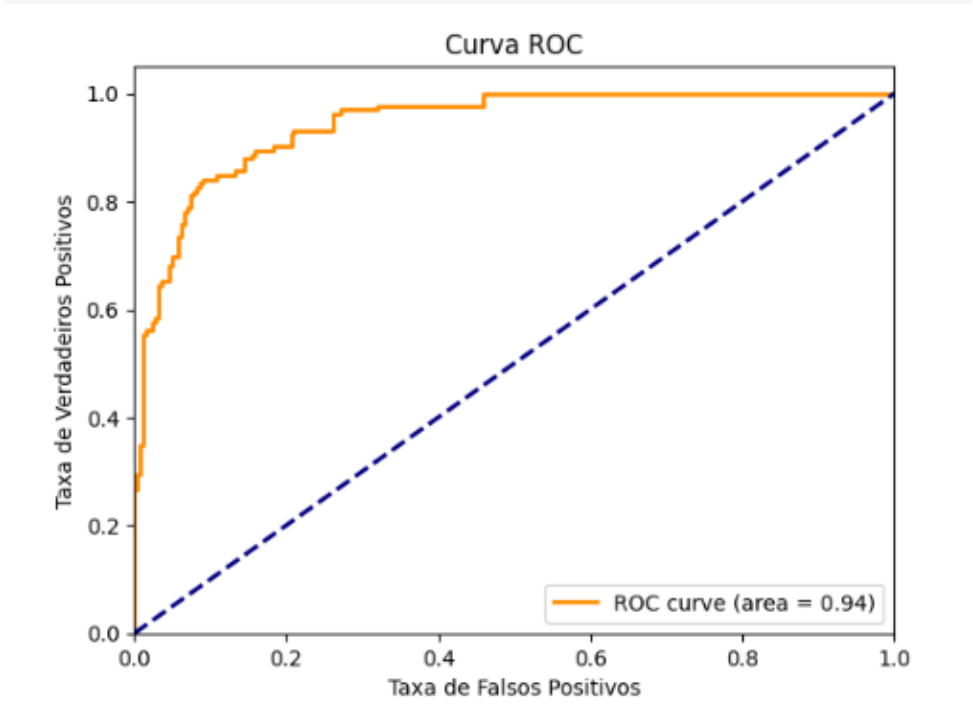
Com uma acurácia de 0.884 e métricas de precisão, recall e f1-score competitivas para ambas as classes, esses resultados reforçam a eficácia da abordagem de classificação na predição da SM com base nas características consideradas neste estudo.

Além disso, foi gerada uma curva ROC para avaliar o desempenho do modelo. Essa curva representa a taxa de FPR em relação à taxa de TPR. A análise dessa curva permite verificar a capacidade do modelo em distinguir corretamente as classes. Ao calcular a AUC, obteve-se um valor de 0.94, indicando uma boa capacidade de classificação do modelo.

A FIGURA 14 apresentada ilustra a curva ROC gerada, que demonstra visualmente a relação entre FPR e TPR. Essa figura complementa os resultados obtidos e reforça a capacidade do modelo de classificação RF em diagnosticar a SM com base nas características analisadas.

Esses resultados indicam que o modelo RF é uma ferramenta promissora para auxiliar no diagnóstico e na predição da SM, proporcionando uma abordagem eficaz e precisa para a predição dessa condição clínica.

Figura 14 – Curva ROC do Modelo RF (Floresta Aleatória) com Marcadores de Framingham



Fonte: Elaboração própria (2023)

Ao traçar a curva ROC foi possível visualizar a relação entre a taxa de falsos positivos e a taxa de verdadeiros positivos. A área AUC foi calculada em 0.94, indicando um desempenho geralmente bom do modelo na classificação das amostras.

Em conclusão, o modelo de classificação RF apresentou um desempenho satisfatório no diagnóstico da SM, com métricas de desempenho competitivas e uma curva ROC que demonstrou

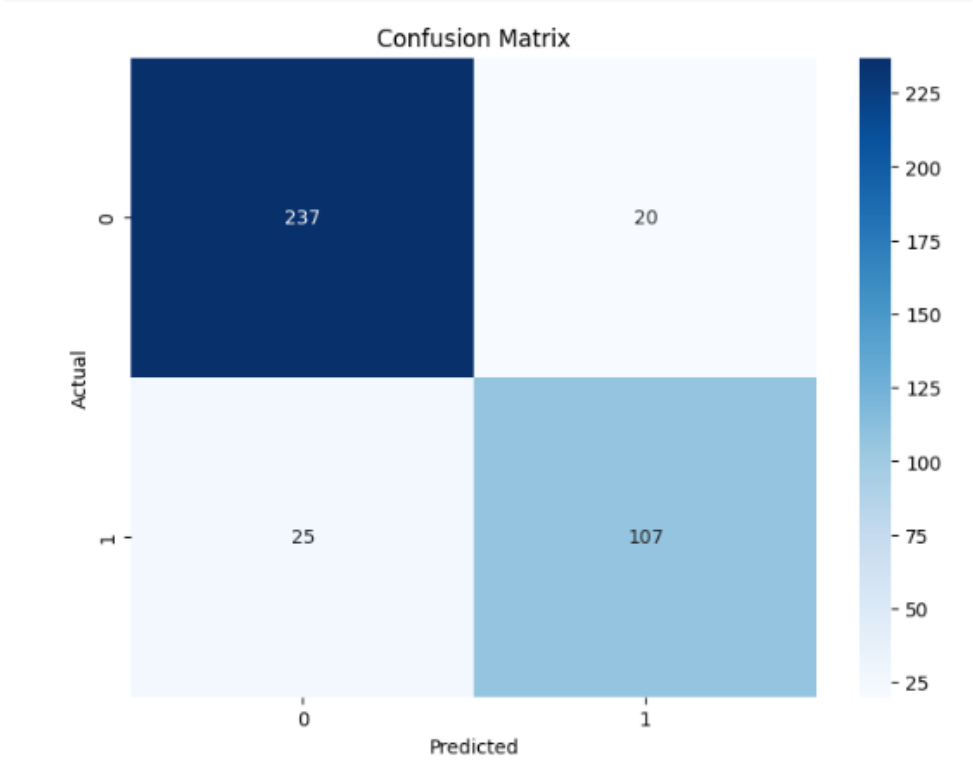
alta capacidade de distinção entre as classes. Esses resultados reforçam a eficácia da abordagem de classificação utilizada para a predição da SM com base nas características consideradas neste estudo. Para complementar a avaliação do modelo, também foi gerada a matriz de confusão, uma representação tabular que permite analisar o desempenho do modelo de classificação. A FIGURA 15 apresenta a representação gráfica da matriz de confusão.

Na interpretação da matriz de confusão, os seguintes resultados são observados:

Na classe 0 (ausência da SM), o modelo classificou corretamente 237 amostras e erroneamente 20 amostras.

Na classe 1 (presença da SM), o modelo classificou corretamente 107 amostras e erroneamente 25 amostras. A matriz de confusão fornece uma visão geral do desempenho do modelo em termos de acertos e erros para cada classe. Essas informações são úteis para avaliar a capacidade do modelo de distinguir corretamente entre as classes e identificar possíveis erros de classificação.

Figura 15 – Matriz de Confusão do Modelo RF (Floresta Aleatória) Utilizando Marcadores de Framingham



Fonte: Elaboração própria (2023)

7.3.2 Resultados e Avaliação dos Modelos de Regressão e Classificação no Cenário 2: Predição da SM utilizando o RFS

No cenário 2, foram explorados os modelos de regressão e classificação para a análise dos marcadores de Framingham na predição da SM. O modelo de regressão RF apresentou um desempenho satisfatório, sendo capaz de prever com precisão a variável dependente contínua associada à SM. Por outro lado, o modelo de classificação RF demonstrou ser a melhor abordagem para o diagnóstico e predição da presença ou ausência da SM, apresentando uma acurácia de 0.884 e métricas de precisão, recall e f1-score competitivas para ambas as classes.

Esses resultados confirmam a eficácia dos modelos de regressão e classificação no contexto da SM, fornecendo informações valiosas para a identificação e predição dessa condição clínica. A análise da importância das características, a curva ROC e a matriz de confusão complementaram a avaliação dos modelos, fornecendo uma visão mais abrangente do desempenho e capacidade discriminativa dos mesmos.

Em suma, o cenário 2 evidencia a utilidade dos modelos de regressão e classificação, destacando o papel dos marcadores de Framingham na predição da SM. Esses resultados têm implicações significativas para a prática clínica, permitindo uma abordagem mais precisa e efetiva no diagnóstico e prevenção dessa condição.

7.4 ANÁLISE E RESULTADOS DO CENÁRIO 3: SELEÇÃO DO MELHOR MODELO DE ML UTILIZANDO OS MARCADORES DA SM

No cenário 3, a apresentação dos resultados segue a mesma abordagem adotada no cenário anterior. Além disso, foram empregados os mesmos procedimentos utilizados no cenário 2, com a distinção de que foram consideradas apenas as variáveis relacionadas à SM, conforme definido pelo critério NCEP-ATP III. Esse critério foi desenvolvido por um painel de especialistas em saúde dos Estados Unidos com o objetivo de diagnosticar e tratar dislipidemias e outras condições associadas à saúde cardiovascular. É relevante ressaltar que a Organização Mundial da Saúde (OMS) e outras instituições internacionais adotam critérios diagnósticos semelhantes para a SM, os quais englobam a presença de pelo menos três dos componentes listados na Tabela 5 de Critérios Diagnósticos para a SM, conforme estabelecido pelo NCEP-ATP III.

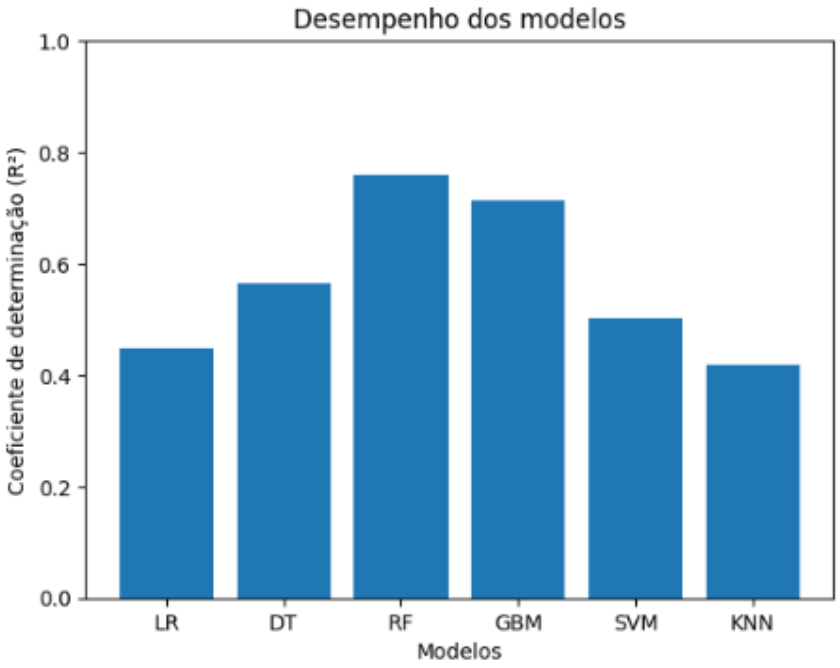
7.4.1 Seleção do Melhor Modelo de Regressão

A seleção do melhor modelo de regressão foi realizada seguindo o procedimento descrito no Apêndice, na seção "Código: Experimentos no Cenário 2 - Regressão". Para realizar os

experimentos com os marcadores específicos, as variáveis relevantes foram selecionadas a partir do DataFrame original. O processo de seleção consiste em criar um novo DataFrame contendo apenas as variáveis desejadas.

No contexto do cenário 3, a seleção cuidadosa das variáveis relevantes é fundamental para a obtenção de um modelo de aprendizado de máquina eficaz. Através desse processo, é possível concentrar a análise nas variáveis mais relevantes para a SM, contribuindo para resultados mais precisos e interpretações mais robustas. A FIGURA 16 apresenta o gráfico que detalha o melhor modelo obtido.

Figura 16 – Desempenho dos Modelos de Regressão Utilizando Marcadores da SM



Fonte: Elaboração própria (2023)

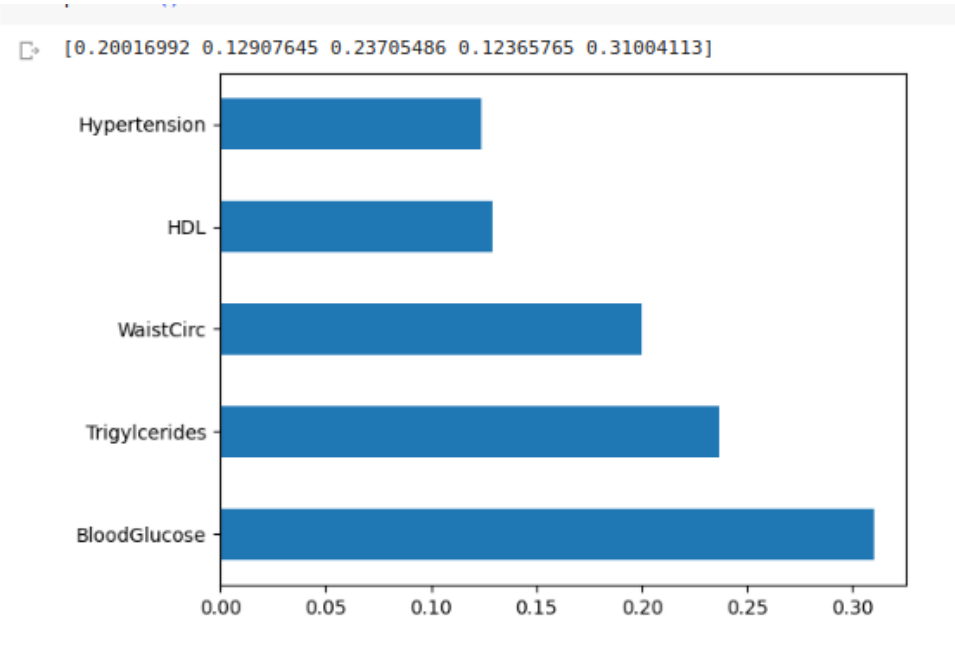
A análise das importâncias das características desempenha um papel fundamental na compreensão da contribuição de cada variável independente no processo de previsão. Essa análise permite visualizar de forma clara e objetiva a influência de cada variável na variável dependente, proporcionando *insights* valiosos sobre quais características têm maior poder de explicação e influência no modelo.

A interpretação dos resultados obtidos pelo modelo de aprendizado de máquina é enriquecida pela análise das importâncias das características. Com base nessa análise, é possível compreender quais variáveis são mais relevantes para a previsão e, assim, tomar decisões embasadas nos resultados.

A representação gráfica das importâncias das características é exibida utilizando a função

plt.show(), como ilustrado na FIGURA 17. Essa visualização fornece uma representação visual do impacto de cada variável e auxilia na identificação das principais influências no processo de previsão.

Figura 17 – Gráfico de Importância das Características de Regressão Utilizando Marcadores de Framingha



Fonte: Elaboração própria (2023)

Utilizando as bibliotecas tabulate e pandas, os resultados foram apresentados de forma organizada na Tabela 7. O modelo selecionado para avaliação foi o RF.

Tabela 7 – Resultados dos modelos de regressão

Modelo	**R2 Score**	**Mean Absolute Error**	**Mean Squared Error**	**RMSE**
RF	0.76	0.10	0.05	0.23

Essas métricas permitem avaliar a qualidade do modelo em relação à sua capacidade de explicar a variação dos dados observados. O valor do R2 Score indica a proporção da variabilidade dos dados que pode ser explicada pelo modelo. Quanto mais próximo de 1, melhor é o desempenho do modelo.

O erro absoluto médio, o erro quadrático médio e a raiz quadrada do erro quadrático médio fornecem medidas quantitativas do desvio entre os valores previstos pelo modelo e os valores reais. Valores mais baixos indicam um melhor ajuste do modelo aos dados observados.

Os resultados apresentados na Tabela 7 demonstram que o modelo RF obteve um desempenho satisfatório de acordo com as métricas utilizadas, evidenciando sua capacidade de realizar

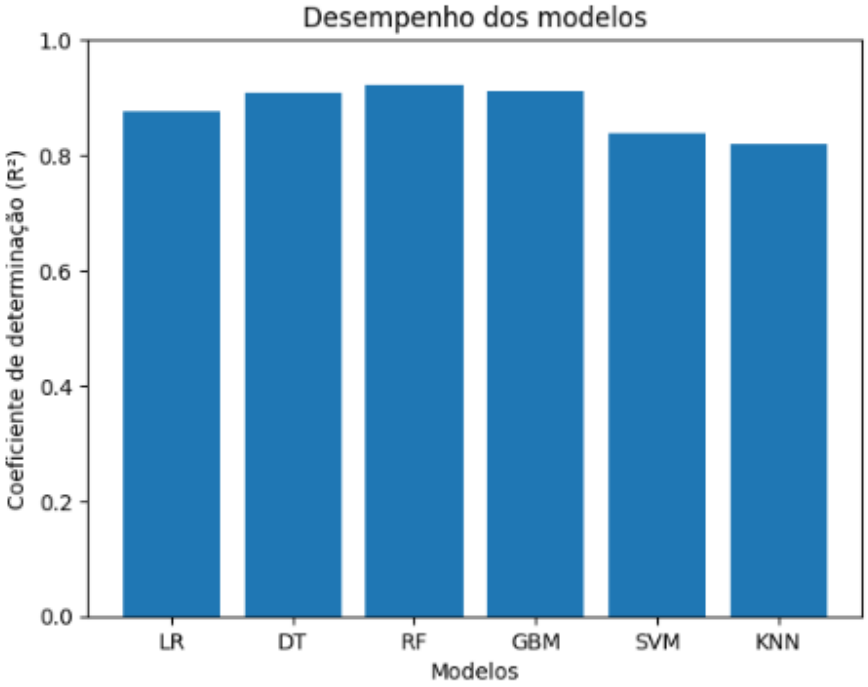
previsões precisas. Essas informações são relevantes para validar a eficácia do modelo e fornecer descobertas para futuras aplicações e melhorias.

7.4.2 Seleção do Melhor Modelo de Classificação

A seleção do melhor modelo de classificação foi conduzida de acordo com o procedimento descrito no Apêndice, na seção intitulada "Código: Experimentos no Cenário 2 - Classificação". Os detalhes sobre a implementação dos modelos de classificação podem ser encontrados minuciosamente no código fornecido no Apêndice A.

Para realizar os experimentos com os marcadores específicos, foram selecionadas as variáveis relevantes a partir do DataFrame original. Esse processo consiste na criação de um novo DataFrame contendo apenas as variáveis desejadas. A seção dedicada ao Cenário 2, focada na seleção do modelo de classificação mais eficaz para o diagnóstico da SM, aborda de forma abrangente a análise dos resultados obtidos. A seguir, serão apresentados os resultados detalhados, fornecendo uma visão aprofundada do desempenho e das capacidades de diagnóstico de cada modelo avaliado. O gráfico destacado na FIGURA 18 ilustra o modelo de classificação mais promissor obtido durante a análise.

Figura 18 – Desempenho dos Modelos de Classificação Utilizando Marcadores da SM



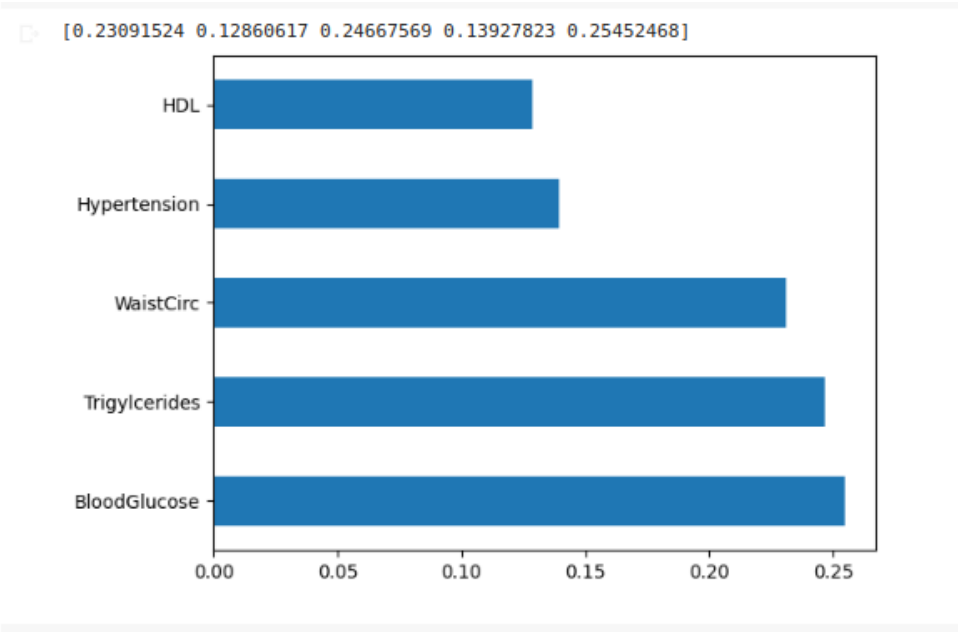
Fonte: Elaboração própria (2023)

Em seguida, foram realizadas a seleção das características com as maiores importâncias e a geração de um gráfico de barras horizontais para representar essas importâncias no contexto

do melhor modelo encontrado. Esse gráfico visualiza de forma intuitiva e concisa as principais variáveis que contribuem para a previsão do diagnóstico da SM.

A FIGURA 19 apresenta o gráfico resultante, proporcionando uma representação visual clara e concisa das importâncias das características nos modelos de classificação. Essa análise é de suma importância para compreender as variáveis que exercem maior influência na classificação da SM, contribuindo para a interpretação dos resultados obtidos e fornecendo iluminações relevantes para futuras investigações nesse campo de estudo.

Figura 19 – Gráfico de Importância das Características de Classificação Utilizando Marcadores da SM



Fonte: Elaboração própria (2023)

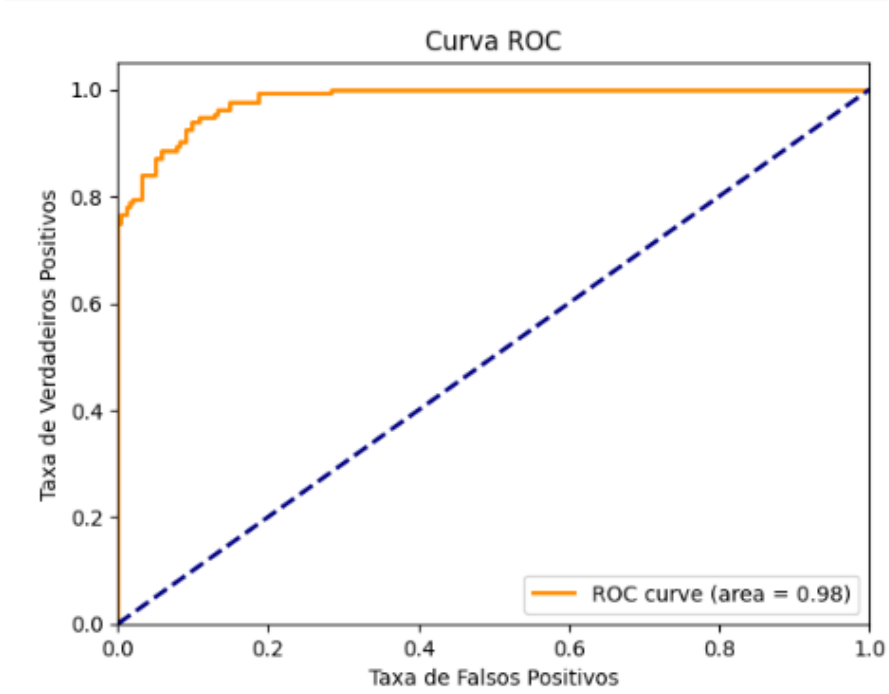
O melhor modelo de classificação, RF, foi avaliado utilizando dados de teste e obteve uma acurácia de 91.26. Isso significa que o modelo foi capaz de realizar previsões corretas em uma taxa satisfatoriamente alta. Os parâmetros ótimos do modelo RF foram identificados como 'max-depth': 7 e 'n-estimators': 500, que controlam a profundidade das árvores de decisão e o número de estimadores no ensemble, respectivamente. Esses parâmetros desempenham um papel crucial na performance do modelo e foram selecionados para maximizar a sua precisão.

Ao analisar o relatório de classificação, é possível observar que o modelo RF obteve resultados promissores para ambas as classes. Para a classe que representa a ausência da SM, o modelo alcançou uma precisão de 0.92, recall de 0.95 e f1-score de 0.93. Já para a classe que representa a presença da SM, a precisão foi de 0.90, recall de 0.84 e f1-score de 0.87. Essas métricas indicam uma capacidade satisfatória do modelo em fazer previsões corretas e identificar corretamente os casos positivos.

Em resumo, o modelo RF otimizado demonstrou um desempenho sólido na tarefa de classificação da SM. Sua alta acurácia, juntamente com as métricas de precisão, recall e f1-score competitivas, destaca sua capacidade de fazer previsões precisas e identificar corretamente os casos positivos. Esses resultados têm implicações importantes para a prática clínica, fornecendo uma abordagem mais precisa e efetiva no diagnóstico e tratamento dessa condição clínica, contribuindo para uma melhor saúde cardiovascular dos pacientes.

Além disso, para complementar a avaliação do modelo, foram geradas duas representações visuais: a curva ROC e a matriz de confusão. A AUC apresentada na FIGURA 20 ilustra a relação entre a taxa de FPR e a taxa de TPR. Essa curva fornece uma visão geral do desempenho do modelo em diferentes pontos de corte e reforça sua capacidade de diagnosticar a SM com base nas características analisadas.

Figura 20 – Curva ROC do Modelo RF (Floresta Aleatória) com Marcadores da SM



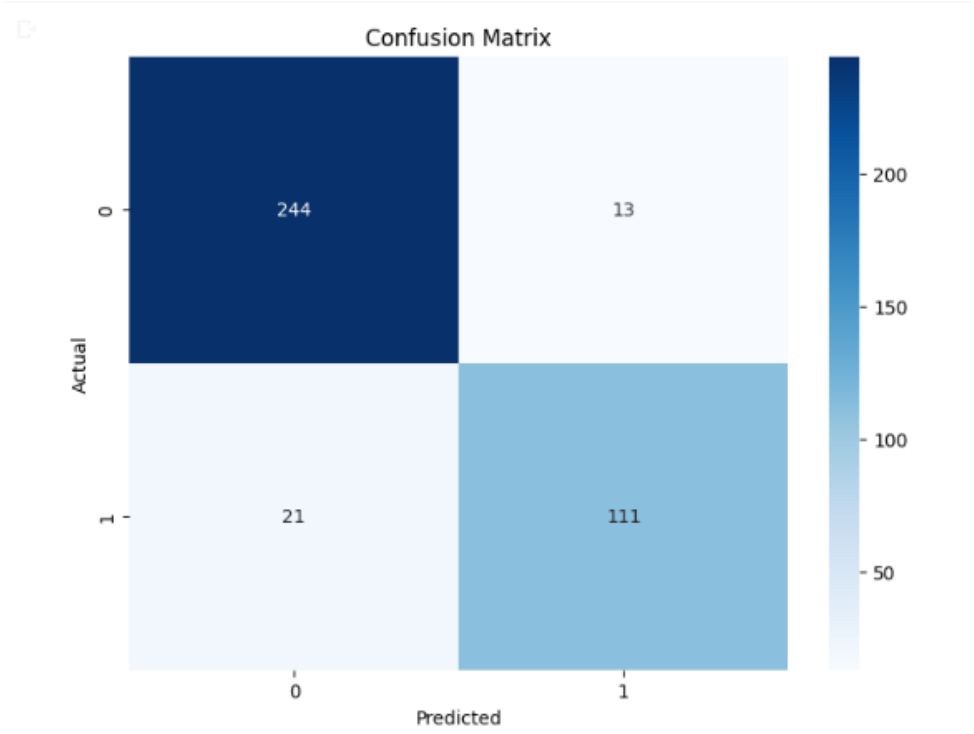
Fonte: Elaboração própria (2023)

Por fim, a matriz de confusão, representada graficamente na FIGURA 21, é uma ferramenta que permite analisar o desempenho do modelo de classificação de forma mais detalhada. Ela mostra a contagem dos casos classificados corretamente e erroneamente para cada classe. Essa representação tabular auxilia na identificação de possíveis erros de classificação e fornece informações valiosas sobre a capacidade discriminativa do modelo em relação à SM.

Em conjunto, a curva ROC e a matriz de confusão complementam os resultados obtidos, reforçando a eficácia do modelo RF na classificação da SM. Essas visualizações fornecem uma

compreensão mais abrangente do desempenho do modelo e podem auxiliar na interpretação dos resultados, bem como na tomada de decisões clínicas para o diagnóstico e tratamento dessa condição clínica.

Figura 21 – Matriz de Confusão do Modelo RF (Floresta Aleatória) Utilizando Marcadores da SM



Fonte: Elaboração própria (2023)

7.5 ANÁLISE E RESULTADOS DO CENÁRIO 4: COMPARAÇÃO E AVALIAÇÃO DOS MODELOS DE REGRESSÃO E CLASSIFICAÇÃO NOS CENÁRIOS 2 E 3: ESCOLHA DO MELHOR MODELO PARA A SM E RDC

O estudo teve como objetivo comparar e avaliar os modelos de regressão e classificação para o desenvolvimento da SM e estimativa do RDC com base nos RFS. Os resultados indicaram que o modelo de regressão RF obteve o melhor desempenho, superando os demais modelos de regressão avaliados nos cenários 2 e 3.

Na tarefa de classificação, o modelo RF também foi identificado como mais eficaz na detecção e diagnóstico da SM e RDC. Apresentou uma acurácia elevada e métricas competitivas de precisão, recall e f1-score, superando os outros classificadores considerados.

Ao analisar apenas os marcadores relacionados à SM, o modelo de regressão RF se destacou novamente como a melhor opção. Demonstrou uma maior capacidade de explicar a variabilidade dos dados e estimar a presença da SM.

De maneira semelhante, na tarefa de classificação, o modelo RF apresentou o melhor desempenho na detecção e diagnóstico da SM e RDC, utilizando os marcadores específicos. Mostrou uma alta acurácia e métricas competitivas, superando os outros classificadores avaliados.

Esses resultados corroboram a seleção do modelo RF como a escolha mais adequada tanto para a tarefa de regressão quanto para a classificação da SM e RDC. Sua capacidade de realizar previsões precisas e identificar corretamente os casos positivos e negativos tem implicações significativas para a prática clínica, permitindo uma abordagem mais eficiente no diagnóstico e tratamento dessa condição clínica. Isso contribui para promover uma melhor saúde cardiovascular nos indivíduos afetados pela SM.

Com base no conhecimento adquirido nos experimentos realizados nos três primeiros cenários, especialmente em relação aos marcadores utilizados no diagnóstico da SM e no FRS, é possível recomendar um modelo de aprendizado de máquina para prever o risco de desenvolvimento de DCV na população ou em indivíduos diagnosticados com SM. No cenário quatro, daremos continuidade aos experimentos com esse conjunto de dados, focalizando agora na análise do mapa de calor e na relação entre as variáveis para encontrar a melhor combinação de marcadores. O objetivo é não apenas recomendar o melhor modelo, mas também identificar quais marcadores dos dois preditores estudados são mais aplicáveis para os propósitos desta pesquisa. No entanto, para validar e aprimorar o modelo, será necessário contar com a cooperação e colaboração de uma equipe multidisciplinar no desenvolvimento de um trabalho futuro.

7.5.1 Resultados e Avaliação dos Modelos de Regressão e Classificação no Cenário 4: Predição da SM e RDC

Nesta seção, apresentamos os resultados e a avaliação dos modelos de regressão e classificação no cenário 4, com o objetivo de prever a ocorrência da SM e o RDC associado a ela, utilizando uma combinação dos marcadores da SM e do FRS. Para essa análise, utilizamos os modelos desenvolvidos nos cenários 2 e 3, a fim de identificar o modelo de Aprendizado de Máquina mais adequado.

As Tabelas 8 e 9 apresentam os resultados dos modelos de regressão, utilizando métricas de desempenho, como o coeficiente de determinação (R^2), e outras métricas relevantes. Optamos por dividir a tabela em duas partes para facilitar a visualização dos resultados. Por sua vez,

as Tabelas 10 e 11 exibem os resultados dos modelos de classificação, considerando diversas métricas de desempenho, incluindo a acurácia (accuracy) e outras relevantes. A divisão em duas partes foi adotada visando a melhor legibilidade e compreensão dos resultados.

A análise desses resultados é de extrema importância, pois fornece informações sobre a eficácia dos modelos na predição da SM e do RDC, com base nos marcadores da SM e do FRS. A partir desses resultados, podemos identificar as combinações de variáveis que apresentaram o melhor desempenho e que podem ser consideradas para a implementação do modelo. Esse modelo pode ser especialmente útil na atenção primária do Sistema Único de Saúde (SUS), onde a disponibilidade de médicos especialistas e recursos financeiros é limitada em certas situações. A capacidade de prever antecipadamente a ocorrência da SM e o RDC pode facilitar a identificação de pacientes que requerem um acompanhamento mais intensivo, encaminhando-os para os cuidados adequados. Isso contribui para a melhoria do atendimento e a otimização dos recursos disponíveis.

7.5.2 Apresentação e Análise dos Resultados dos Modelos de Regressão Treinados e Testados

As Tabelas 8 e 9 apresentam os resultados dos modelos de regressão utilizando diversas métricas de desempenho, incluindo as métricas de desempenho, como o coeficiente de determinação (R^2) e outras relevantes. A divisão em duas partes foi adotada com o objetivo de aprimorar a legibilidade e facilitar a compreensão dos resultados obtidos.

Tabela 8 – Desempenho dos modelos de regressão para Síndrome Metabólica e Risco Cardiovascular - Parte 1

Modelo	Marcadores SM	Marcadores SM + Sexo	Marcadores SM + Idade	Marcadores SM + Obesidade
LR	0.4496	0.4840	0.4037	0.5872
DT	0.5643	0.9083	0.4955	0.8165
RF	0.7608	0.9407	0.7682	0.8928
GBM	0.7130	0.7901	0.7162	0.7836
SVM	0.5022	0.4977	0.5056	0.4982
KNN	0.4193	0.4175	0.4689	0.4212

A descrição das siglas estão disponíveis na lista de siglas

Tabela 9 – Desempenho dos modelos de regressão para Síndrome Metabólica e Risco Cardiovascular - Parte 2

Modelo	Marcadores SM + FRS	Marcadores FRS	Marcadores FRS + Obesidade	Marcadores FRS + SM + Obesidade
LR	0.4381	0.1859	0.4267	0.5643
DT	0.9197	0.3120	0.5413	0.9312
RF	0.9304	0.6016	0.7328	0.9359
GBM	0.7855	0.5832	0.6753	0.8022
SVM	0.5000	0.3350	0.3497	0.4988
KNN	0.4693	0.3680	0.3652	0.4693

As tabelas 8 e 9 fornecem uma análise que permite identificar o modelo mais eficiente e a combinação mais simples de marcadores para a predição da SM. Ao avaliar as métricas de desempenho, como o coeficiente de determinação (R^2), constatou-se que o modelo RF apresentou resultados consistentes em todas as combinações de marcadores. Dentre essas combinações, destaca-se aquela que incorpora os marcadores da SM em conjunto com o marcador de sexo, evidenciando um coeficiente de determinação de 0.94.

A união dos marcadores da SM e do sexo demonstrou um desempenho superior em termos de precisão na previsão da SM, sugerindo que o sexo do indivíduo pode ser um fator relevante na identificação e prognóstico dessa condição de saúde.

Portanto, é recomendável adotar o RF em conjunto com a combinação de marcadores da SM e sexo como a abordagem mais simples e eficaz para a previsão da SM na atenção primária do SUS. Essa combinação de marcadores pode fornecer informações valiosas para identificar indivíduos com maior risco de desenvolver a síndrome e encaminhá-los para um acompanhamento mais adequado e personalizado.

7.5.3 Apresentação e Análise dos Resultados dos Modelos de Classificação Treinados e Testados

As Tabelas 10 e 11 exibem os resultados dos modelos de classificação com base em diversas métricas de desempenho, incluindo a acurácia (accuracy) e outras relevantes. A divisão em duas partes foi adotada para melhorar a legibilidade e facilitar a compreensão dos resultados.

Tabela 10 – Desempenho dos modelos de classificação para Síndrome Metabólica e Risco Cardiovascular - Parte 1

Modelo	Marcadores SM	Marcadores SM + Sexo	Marcadores SM + Idade	Marcadores SM + Obesidade
LR	0.8766	0.8843	0.8663	0.9074
DT	0.9074	0.9794	0.8946	0.9640
RF	0.9228	0.9742	0.9074	0.9562
GBM	0.9100	0.9922	0.9177	0.9665
SVM	0.8380	0.8380	0.8508	0.8380
KNN	0.8200	0.8200	0.8354	0.8226

A descrição das siglas estão disponíveis na lista de siglas

Tabela 11 – Desempenho dos modelos de classificação para Síndrome Metabólica e Risco Cardiovascular - Parte 2

Modelo	Marcadores SM + FRS	Marcadores FRS	Marcadores FRS + Obesidade	Marcadores FRS + SM + Obesidade
LR	0.8637	0.8174	0.8714	0.9023
DT	0.9820	0.8303	0.8868	0.9871
RF	0.9691	0.8817	0.9048	0.9794
GBM	0.9948	0.8740	0.9125	0.9974
SVM	0.8483	0.7917	0.7969	0.8457
KNN	0.8354	0.7969	0.7969	0.8354

Ao analisar os resultados das Tabelas 10 e 11, que exibem o desempenho dos modelos de classificação para a SM e o RDC, observam-se os seguintes aspectos relevantes:

Na Tabela 10 - Parte 1:

O modelo RF obteve o melhor desempenho, apresentando valores de classificação variando entre 0.9228 e 0.9742, dependendo dos marcadores utilizados. Os DT e GBM também apresentaram resultados positivos, com valores de classificação superiores a 0.9 em diversas configurações de marcadores. O modelo de LR demonstrou consistência nos resultados, com valores de classificação acima de 0.86 em todas as configurações.

Na Tabela 11 - Parte 2:

O modelo GBM exibiu o melhor desempenho, com valores de classificação variando entre 0.9948 e 0.9974, dependendo dos marcadores utilizados. Similarmente à Tabela 10, o modelo RF também demonstrou resultados sólidos, com valores de classificação acima de 0.96 em todas as configurações. Os modelos LR, DT, SVM e KNN apresentaram um desempenho razoável, com valores de classificação variando entre 0.7917 e 0.9023, dependendo dos marcadores utilizados.

Em resumo, os modelos RF, GBM e LR evidenciaram um bom desempenho geral em ambas as tabelas, com valores de classificação consistentes. Esses modelos podem ser considerados como os principais candidatos para o diagnóstico da SM e do RDC, com base nos resultados apresentados. No entanto, vale ressaltar que a escolha do melhor modelo também depende dos critérios específicos do problema e das características dos dados. É relevante destacar que o GBM obteve valores de classificação excepcionalmente elevados, variando entre 0.9948 e 0.9974, na Tabela 11 - Parte 2, indicando um desempenho notável na previsão da SM e do RDC. Ademais, o GBM também apresentou resultados positivos na Tabela 10 - Parte 1, com valores classificação acima de 0.91 em todas as configurações.

No entanto, essa sugestão se baseia, exclusivamente, nos resultados obtidos nos experimentos realizados nos cenários 2, 3 e 4. Para uma seleção mais precisa e confiável, recomenda-se realizar uma análise mais aprofundada, levando em consideração outros fatores relevantes, como a interpretabilidade do modelo, a complexidade computacional, a disponibilidade de recursos e os requisitos específicos do problema em questão.

7.6 ESCOLHA DO MELHOR MODELO DE REGRESSÃO E CLASSIFICAÇÃO DO RDC ASSOCIADO A SMA

A fim de identificar o modelo mais adequado e confiável para a predição da SM e DRC, foi realizada uma análise minuciosa dos resultados dos modelos de regressão e classificação. Essa análise considerou as variáveis mais importantes e relevantes do dataset, relacionadas ao risco de desenvolvimento dessas condições de saúde. O modelo RF destacou-se como o mais eficaz e confiável na predição da SM. Essa conclusão foi reforçada pelos resultados obtidos nos cenários 2 e 3, onde o RF superou os demais modelos avaliados de forma consistente.

No contexto dos modelos de regressão, o RF apresentou resultados consistentes em todas as combinações de variáveis. Notadamente, a combinação que envolveu os marcadores da SM juntamente com o marcador de sexo demonstrou um coeficiente de determinação (R^2) de 0.94, indicando um bom ajuste do modelo aos dados.

No que diz respeito aos modelos de classificação, o RF também se destacou, alcançando altos índices de acurácia em todas as combinações de variáveis. Especificamente, a combinação dos marcadores da SM com o marcador de sexo obteve uma acurácia de 0.974, evidenciando a capacidade do modelo em classificar corretamente os indivíduos em relação à presença ou ausência da síndrome.

Com base nessas análises, é recomendado o uso do modelo RF tanto para a regressão

quanto para a classificação da SM. A combinação de marcadores da SM com o marcador de sexo demonstrou-se ser a abordagem mais simples e eficaz para a predição dessas condições de saúde.

A utilização desse modelo e combinação de variáveis pode fornecer informações valiosas na identificação de indivíduos com maior risco de desenvolver a síndrome, permitindo intervenções e acompanhamento adequados e personalizados.

É fundamental salientar que essas conclusões são específicas para o presente estudo e é recomendada a realização de análises adicionais e validações em estudos futuros, a fim de garantir a robustez e generalização dos resultados obtidos.

7.7 ANÁLISE E RESULTADOS DO CENÁRIO 5: SELEÇÃO DO MELHOR MODELO DE ML USANDO MARCADORES DA SM E O FRS

Neste cenário, seguiremos a mesma metodologia adotada nos cenários anteriores para apresentar os resultados. Inicialmente, serão apresentados os resultados dos experimentos de regressão, com detalhes sobre as métricas de desempenho e interpretação dos resultados. Em seguida, abordaremos os resultados encontrados para o modelo de classificação, realizando uma análise similar.

7.7.1 Seleção do Melhor Modelo de Regressão

Neste cenário, utilizamos a metodologia adotada nos cenários anteriores para selecionar o melhor modelo de regressão para a predição do RDC em dados combinados de doenças cardíacas. Foram realizados três experimentos com os mesmos modelos e metodologia. No primeiro experimento, todas as variáveis do dataset foram utilizadas para treinar e testar o modelo proposto, a fim de escolher o melhor modelo. Os resultados obtidos estão apresentados na Tabela 12.

No segundo experimento, o modelo foi treinado e testado apenas com as variáveis relacionadas ao diagnóstico da SM, com o objetivo de investigar a relação entre RDC e essa síndrome. Já no terceiro experimento, o modelo foi treinado com cinco variáveis, incluindo os marcadores da SM e o FRS.

O objetivo principal desta pesquisa é recomendar um modelo de aprendizado de máquina capaz de diagnosticar e prever o RDC associado à SM. A compreensão dessa relação é crucial para avançar o conhecimento na área e desenvolver estratégias mais eficazes de prevenção e tratamento. Os resultados dos experimentos de regressão fornecem importantes descobertas sobre as métricas de desempenho e interpretação dos modelos, contribuindo para a melhoria do

cuidado dos pacientes e a promoção da saúde cardiovascular.

Tabela 12 – Seleção do Melhor Modelo de Regressão para Risco Cardiovascular em Dados Combinados de Doenças Cardíacas

Modelo	Todos Marcadores do Dataset	Três Marcadores da SM	Cinco Marcadores SM + FRS
LR	0.34	-0.43	-0.19
DT	0.56	-0.10	-0.02
RF	0.70	0.15	0.34
GBM	0.62	0.15	0.26
SVM	0.22	-0.07	0.11
KNN	0.21	0.06	0.16

Neste cenário, foram realizados os mesmos procedimentos adotados nos cenários anteriores, com foco na análise dos modelos de classificação treinados e testados. A tabela 13 apresenta os resultados obtidos.

Tabela 13 – Seleção do Melhor Modelo de Classificação para Risco Cardiovascular em Dados Combinados de Doenças Cardíacas

Modelo	Todos Marcadores do Dataset	Três Marcadores da SM	Cinco Marcadores SM + FRS
LR	0.72	0.71	0.71
DT	0.89	0.77	0.78
RF	0.95	0.80	0.84
GBM	0.91	0.72	0.72
SVM	0.72	0.66	0.66
KNN	0.71	0.70	0.70

No presente estudo, foi conduzida uma análise para selecionar o modelo de regressão e classificação mais adequado para a predição do RDC no conjunto de dados combinados de doenças cardíacas, contendo pacientes da Cleveland Clinic, Hungria, Suíça e VA Long Beach. Diversas combinações de marcadores e variáveis foram exploradas nesse contexto.

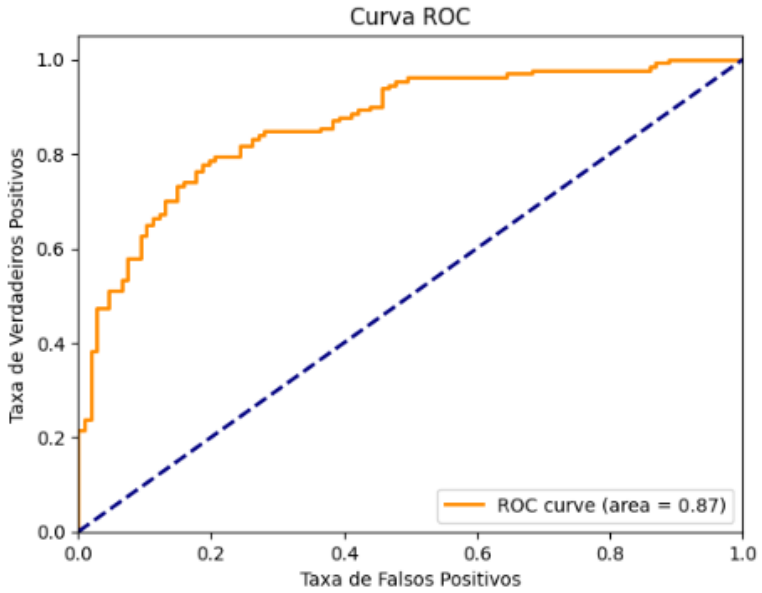
Na análise de regressão, observou-se que o modelo RF apresentou os melhores resultados, evidenciando uma correlação significativa com os dados. Notavelmente, o modelo que incorpora cinco marcadores, combinando variáveis da SM com o FRS (SM + FRS), alcançou uma correlação positiva ($R^2 = 0.341$), revelando uma associação relevante com o RDC. No entanto, é importante destacar que esse valor é consideravelmente inferior ao obtido no teste com todas as variáveis (0.70), indicando que a simplificação do modelo pode levar a uma perda de precisão.

Quanto à análise de classificação, também foi identificado o modelo RF como o de melhor desempenho, considerando as métricas de avaliação. O modelo que utiliza todas as variáveis do conjunto de dados combinado obteve a maior acurácia (0.953), destacando-se como o mais eficaz na classificação do RDC.

Vale ressaltar que tanto o modelo com três marcadores da SM quanto o modelo com cinco marcadores da SM combinados ao FRS (SM + FRS) apresentaram desempenho considerável, com correlações e acurácias significativas (0.80 e 0.84, respectivamente). Essas combinações mais simples de variáveis podem ser uma abordagem interessante, uma vez que fornecem informações relevantes para a predição do RDC de forma menos invasiva, evitando a necessidade de exames mais complexos.

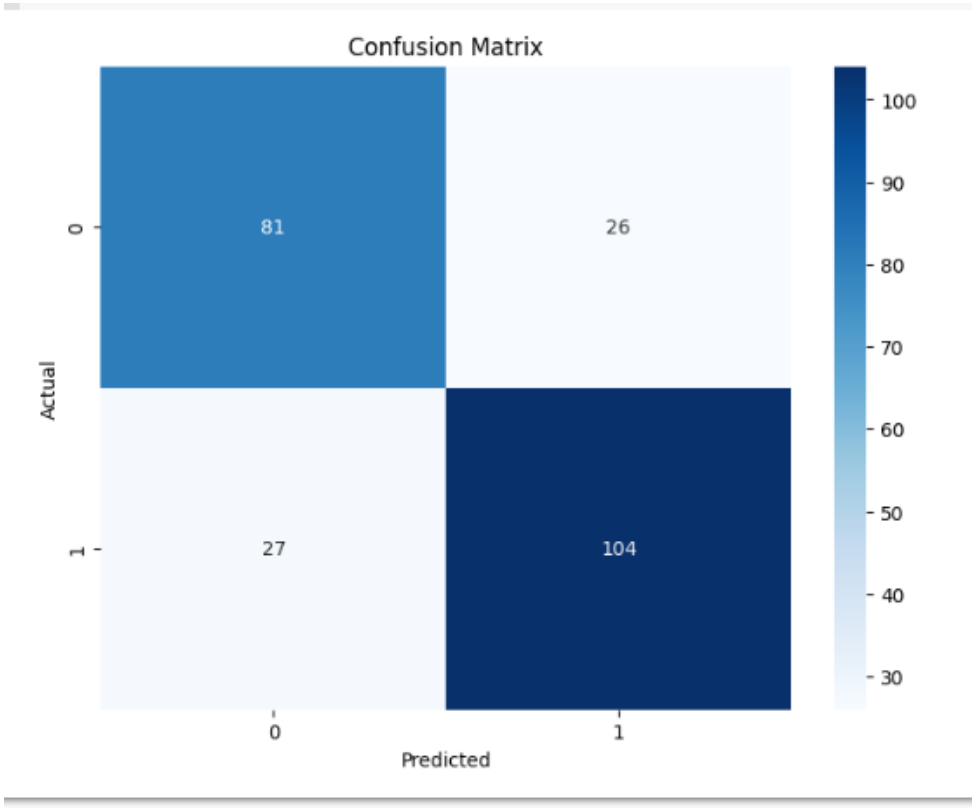
Assim, com base nos resultados obtidos, o modelo RF mostrou-se o mais adequado para a tarefa de regressão e classificação do RDC neste estudo. Além disso, a combinação de variáveis da SM com o FRS apresentou-se promissora, oferecendo uma alternativa mais simples e eficaz para a análise e previsão do RDC. A seguir, serão apresentados os resultados do melhor modelo, que incorpora os marcadores da SM juntamente com os marcadores utilizados no FRS. Será fornecida a Curva ROC na FIGURA 22 e a matriz de confusão na FIGURA 23. Posteriormente, serão discutidos os resultados obtidos nesse modelo, após a seleção do melhor modelo e o ajuste de seus parâmetros, seguido de uma breve explicação.

Figura 22 – Curva ROC do Modelo RF com Marcadores da SM e do FRS



Fonte: Elaboração própria (2023)

Figura 23 – Matriz de Confusão do Modelo RF Utilizando Marcadores da SM e do RFS



Fonte: Elaboração própria (2023)

O modelo Random Forest obteve um resultado promissor na análise do RDC. A curva ROC apresentou uma área de 0.87, o que indica uma boa capacidade de discriminação entre as classes positiva e negativa. Esse valor representa a probabilidade de que o modelo classifique corretamente uma amostra aleatória da classe positiva com uma pontuação de probabilidade mais alta do que uma amostra aleatória da classe negativa.

Ao avaliar a matriz de confusão, podemos observar que o modelo classificou corretamente 81 amostras da classe 0 (RDC negativo) e 104 amostras da classe 1 (RDC positivo). No entanto, houve algumas amostras classificadas erroneamente, com 26 amostras da classe 0 e 27 amostras da classe 1 sendo classificadas incorretamente.

As métricas de precisão, recall e F1-score fornecem informações mais detalhados sobre o desempenho do modelo. Para a classe 0, o modelo obteve uma precisão de 0.75, um recall de 0.76 e um F1-score de 0.75. Esses resultados indicam que o modelo apresentou uma boa capacidade de identificar corretamente os casos negativos de RDC, com um equilíbrio adequado entre precisão e recall.

Para a classe 1, a precisão foi de 0.80, o recall foi de 0.79 e o F1-score foi de 0.80. Isso indica que o modelo também obteve um bom desempenho em identificar corretamente os casos positivos de RDC.

A acurácia geral do modelo foi de 0.78, o que significa que aproximadamente 78% das amostras foram classificadas corretamente. A média ponderada do F1-score foi de 0.78, levando em consideração o suporte (número de amostras) de cada classe.

Além disso, foi realizada a validação cruzada com 5 folds para aprimorar e explorar a validação do melhor modelo escolhido. Essa abordagem contribui para verificar a robustez e a generalização do modelo, garantindo resultados mais confiáveis.

Em resumo, os resultados mostram que o modelo Random Forest apresentou um desempenho satisfatório na classificação do RDC, com uma área significativa sob a curva ROC e métricas sólidas de precisão, recall e F1-score para ambas as classes. No entanto, é importante realizar uma análise mais aprofundada e considerar outros fatores relevantes antes de concluir sobre a eficácia do modelo.

7.8 ANÁLISE E RESULTADOS DO CENÁRIO 6: SELEÇÃO DO MELHOR MODELO DE ML USANDO MARCADORES DA SM E FRS COM O DATASET FRAMINGHAM.CSV

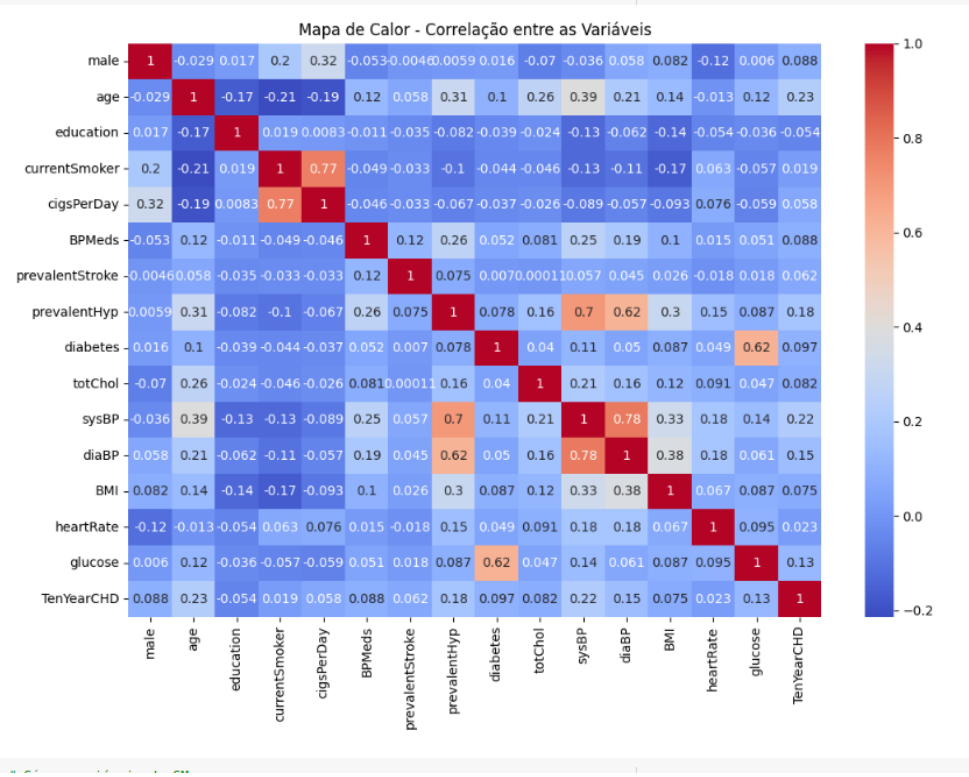
Com o objetivo de recomendar um modelo de aprendizado de máquina para a predição do RDC relacionado à SM, é crucial considerar o cenário do estudo de Framingham, uma vez que o banco de dados utilizado está diretamente relacionado a essa pesquisa.

Para atingir o objetivo de desenvolver um modelo de predição de RDC, foram realizadas as seguintes etapas, conforme descrito na seção de metodologia dos experimentos: importação e visualização dos dados, pré-processamento dos dados (normalização e divisão em conjuntos de treinamento e teste), treinamento de diversos modelos de classificação e regressão, avaliação do desempenho dos modelos por meio de métricas específicas de regressão para os modelos de regressão e métricas específicas de classificação para os modelos de classificação, seleção do melhor modelo e ajuste de seus parâmetros, e, por fim, reavaliação do modelo finalizado.

Na primeira parte, foi realizada a limpeza dos dados e uma análise exploratória. Identificaram-se valores nulos em várias variáveis do conjunto de dados. A coluna "education"apresentou 105 valores nulos, "cigsPerDay"registrou 29 valores nulos, "BPMeds"apresentou 53 valores nulos, "totChol"mostrou 50 valores nulos, "BMI"teve 19 valores nulos, "heartRate"apresentou 1 valor nulo e "glucose"registrou 388 valores nulos. Esses valores ausentes foram tratados antes de realizar a modelagem dos dados. Em seguida, foram calculadas as correlações entre as variáveis, com o intuito de compreender a relação entre elas e identificar possíveis problemas de multicolinearidade. Para isso, utilizou-se a biblioteca seaborn para gerar um mapa de calor (heatmap) das

correlações, conforme apresentado na FIGURA 24.

Figura 24 – Mapa de Calor - Correlação entre as Variáveis



Fonte: Elaboração própria (2023)

Durante a análise de regressão foi constatado que o modelo GradientBoostingRegressor apresentou os melhores resultados, indicando uma correlação significativa com os dados. No entanto, os resultados obtidos para a seleção do melhor modelo de regressão não foram satisfatórios. Por outro lado, nos experimentos de classificação, o modelo Random Forest mostrou-se mais adequado para classificar o RDC em um período de 10 anos, conforme mencionado no RFS (National Heart, Lung, and Blood Institute, 1948).

Os resultados das métricas para o melhor modelo foram os seguintes: A Tabela 14 mostra os resultados obtidos na escolha do melhor Modelo de regressão:

Tabela 14 – Resultados dos modelos de regressão

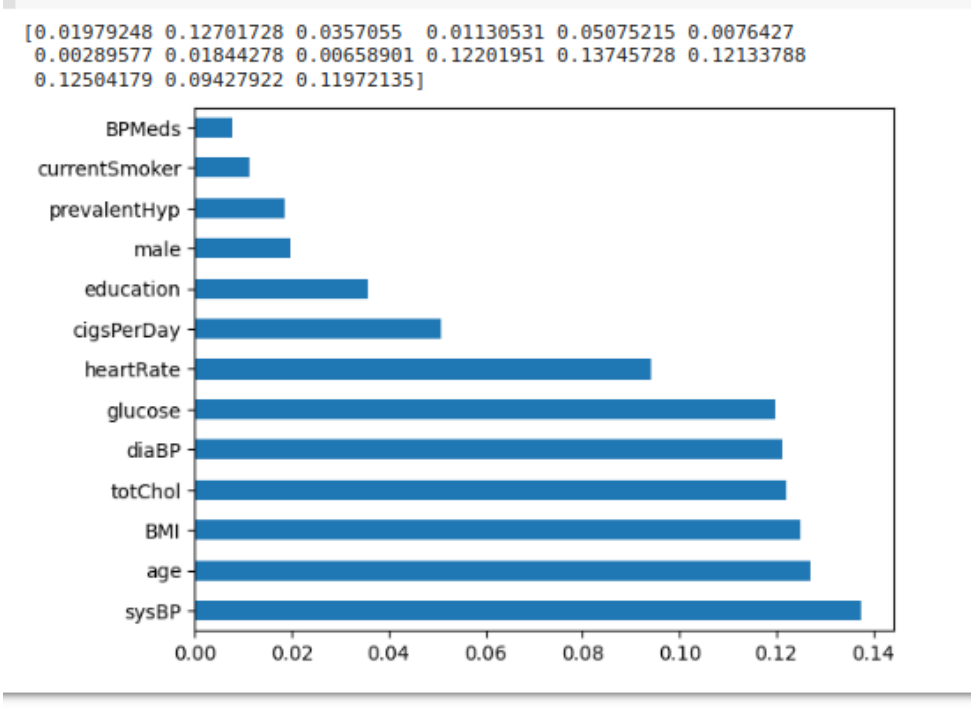
Modelo	**R2 Score**	**Mean Absolute Error**	**Mean Squared Error**	**RMSE**
GBM	0.08	0.23	0.12	0.35

Os resultados obtidos pelos dois modelos testados revelaram uma notável similaridade, destacando-se a diferença mais significativa na matriz de confusão, uma métrica amplamente recomendada para a análise desse tipo de modelo. O primeiro modelo considerou todas as variáveis disponíveis no conjunto de dados, enquanto o segundo se restringiu às variáveis

associadas ao diagnóstico da SM. É válido ressaltar que a acurácia do melhor modelo foi de 0.83. Esses resultados encorajam investigações futuras no sentido de desenvolver calculadoras e/ou escores para a predição do RDC em pacientes diagnosticados com SM, assim como para prever o risco de desenvolvimento dessa síndrome, utilizando critérios diagnósticos específicos da SM e estabelecendo uma relação com o RFS. O emprego de técnicas de aprendizado de máquina, como proposto neste estudo, pode contribuir para atingir esses objetivos.

A FIGURA 25 apresenta o gráfico resultante, proporcionando uma representação visual clara e concisa das importâncias das características nos modelos de classificação. Essa análise é de suma importância para compreender as variáveis que exercem maior influência na classificação do RDC, contribuindo para a interpretação dos resultados obtidos e fornecendo descobertas relevantes para futuras investigações nesse campo de estudo.

Figura 25 – Gráfico de Importância das Características de Classificação Utilizando Todos os Marcadores



Fonte: Elaboração própria (2023)

Os resultados apresentam as métricas de desempenho dos modelos de classificação para a predição do RDC no conjunto de dados de Framingham, considerando todos os marcadores do dataset e apenas os marcadores relacionados à SM.

Os modelos LR, GBM, SVM e KNN obtiveram acurácia de 0.833 em ambas as configurações de marcadores. Isso indica que esses modelos tiveram um desempenho consistente na predição do RDC, independentemente dos marcadores considerados.

Tabela 15 – Melhor Modelo de Classificação para Risco Cardiovascular no conjunto de dados de Framingham

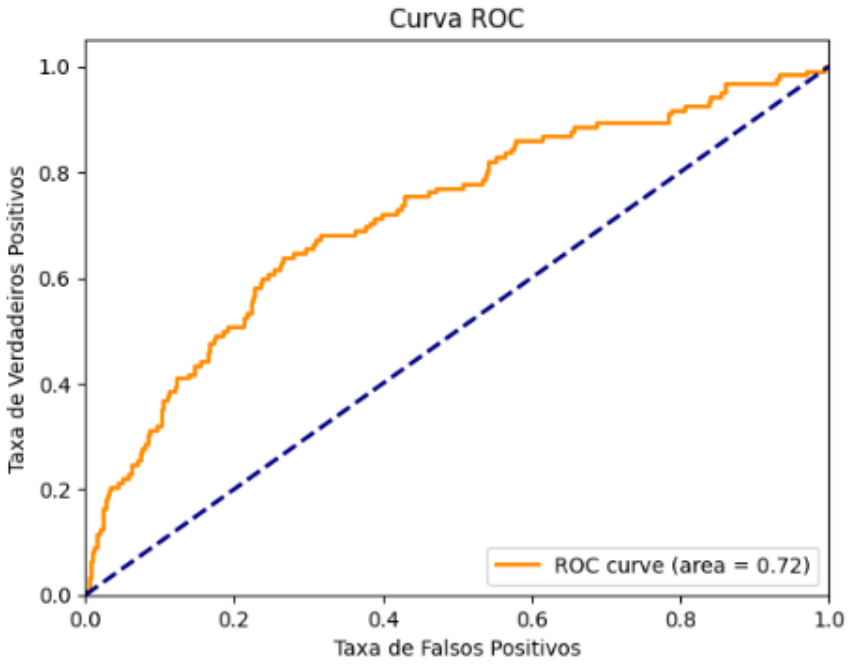
Modelo	Todos Marcadores do Dataset	Só Marcadores da SM
LR	0.8333	0.8333
DT	0.7322	0.7322
RF	0.8346	0.8346
GBM	0.8333	0.8333
SVM	0.8333	0.8333
KNN	0.8265	0.8265

Por outro lado, o modelo DT apresentou uma acurácia ligeiramente inferior, com valor de 0.732, enquanto o modelo RF obteve uma acurácia mais elevada, alcançando 0.835 em ambas as configurações. Esses resultados sugerem que o modelo RF pode ser mais eficaz na predição do RDC no conjunto de dados de Framingham.

Com base nos resultados obtidos, o modelo RF demonstrou ser o mais adequado para a tarefa de classificação do RDC neste estudo. A Tabela 15 apresenta os resultados de acurácia dos modelos obtidos.

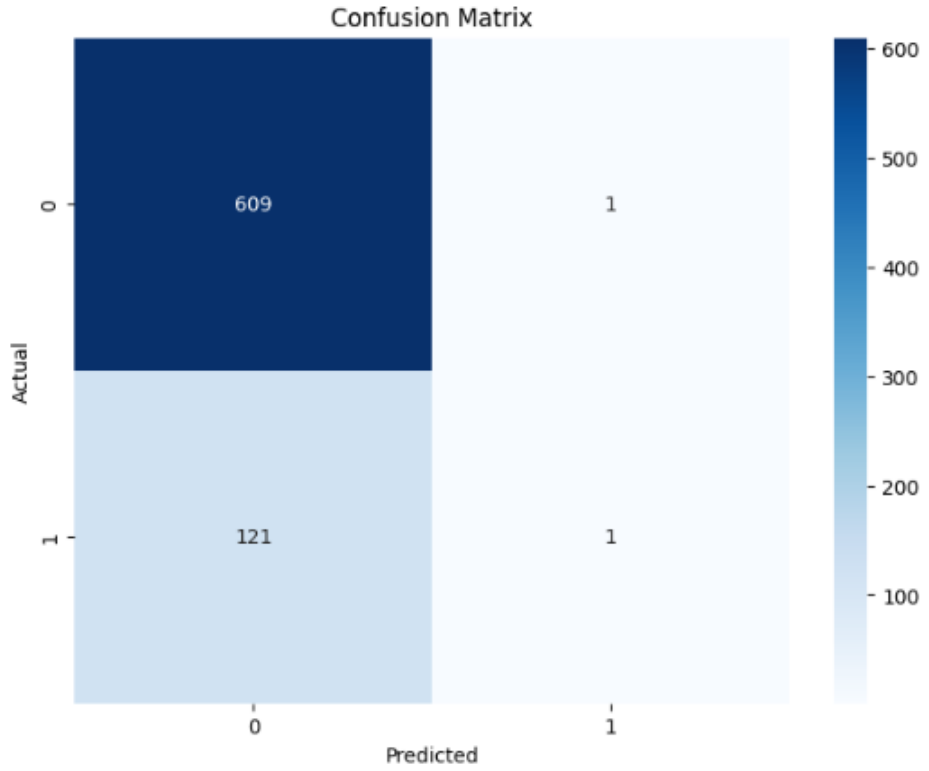
Além das métricas de desempenho, serão fornecidas visualizações adicionais, como a Curva ROC na FIGURA 26 e a matriz de confusão na FIGURA 27. Essas visualizações ajudarão a entender melhor os modelos treinados e testados, oferecendo revelações para uma análise mais abrangente.

Figura 26 – Curva ROC do Modelo RF com Marcadores da SM e do FRS



Fonte: Elaboração própria (2023)

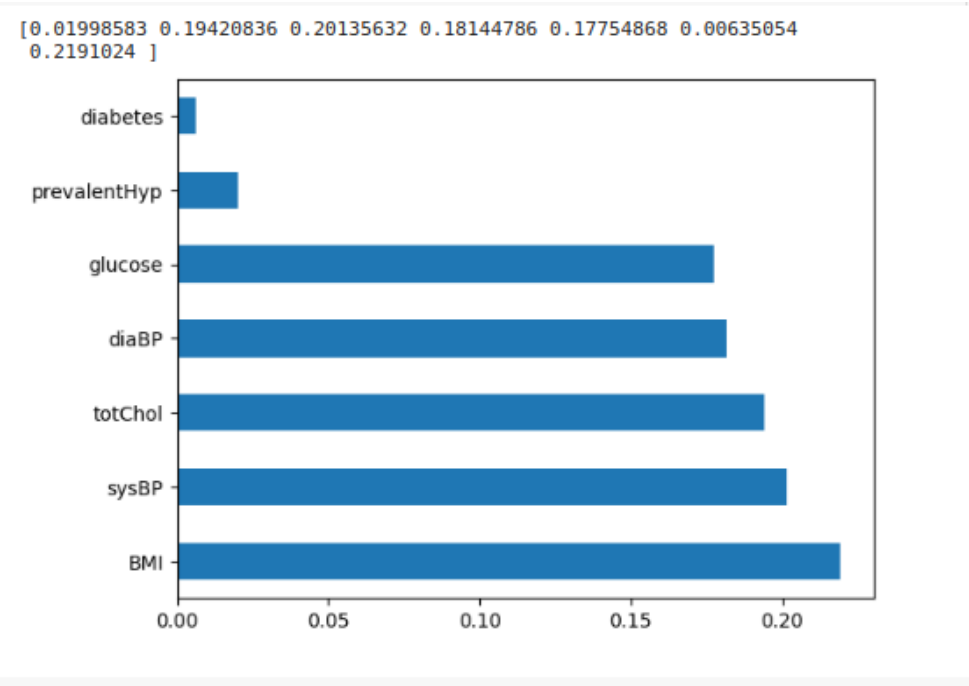
Figura 27 – Matriz de Confusão do Modelo RF Utilizando Marcadores da SM e do FRS



Fonte: Elaboração própria (2023)

Em seguida, realizou-se uma análise dos modelos treinados considerando apenas as variáveis relacionadas ao diagnóstico da SM. A FIGURA 28 exibe o gráfico resultante, fornecendo uma representação visual clara e concisa da importância das características nos modelos de classificação.

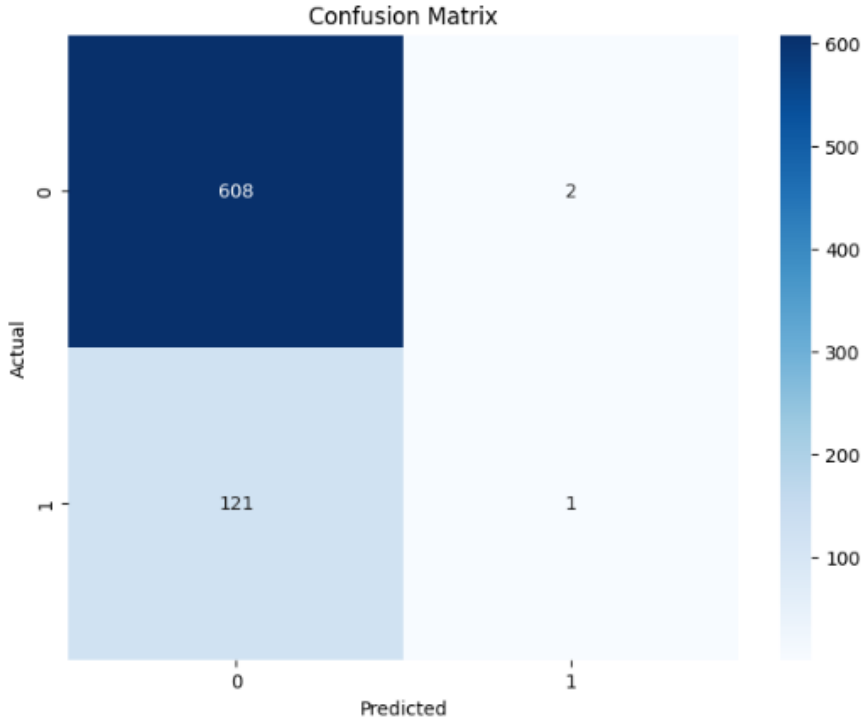
Figura 28 – Gráfico de Importância das Características de Classificação Utilizando Marcadores da SM



Fonte: Elaboração própria (2023)

Como mencionado anteriormente, a curva ROC resultou em um desempenho idêntico para os dois modelos testados. Portanto, na FIGURA 29, será apresentada a Matriz de Confusão para analisar a diferença entre esses modelos, levando em consideração todos os passos recomendados na metodologia deste estudo.

Figura 29 – Matriz de Confusão do Modelo RF Utilizando Marcadores da SM e do FRS



Fonte: Elaboração própria (2023)

Os resultados da matriz de confusão revelaram informações significativas sobre o desempenho dos modelos de classificação no contexto da predição do RDC no conjunto de dados de Framingham. O primeiro modelo, que considerou todas as variáveis disponíveis, obteve uma taxa de acerto elevada para a classe negativa, indicando a capacidade de identificar corretamente os casos não pertencentes à classe de RDC. No entanto, apresentou dificuldades na identificação dos casos positivos, com um número considerável de falsos negativos e falsos positivos.

Por outro lado, o segundo modelo, que utilizou apenas as variáveis relacionadas à SM, apresentou uma taxa de acerto ainda maior para a classe negativa. Isso sugere que essas variáveis têm um impacto significativo na predição do RDC. No entanto, a taxa de acerto para a classe positiva foi muito baixa, com apenas uma previsão correta. Isso ressalta a necessidade de melhorias e ajustes nos modelos para aprimorar a capacidade de identificar corretamente os casos de RDC.

A análise comparativa entre os dois modelos revelou que, apesar das diferenças na matriz de confusão, ambos obtiveram resultados semelhantes em termos de acurácia geral. Isso indica que a inclusão das variáveis relacionadas à SM não resultou em uma melhoria significativa no desempenho da predição em comparação com o uso de todas as variáveis disponíveis.

Esses resultados são de grande relevância, pois fornecem informações valiosas para

pesquisadores e profissionais da área de saúde. Eles destacam a importância das variáveis relacionadas à SM na predição do RDC e ressaltam a necessidade de aprimoramentos nos modelos para melhor identificação dos casos positivos. Futuras pesquisas podem explorar outras abordagens de modelagem e considerar a inclusão de mais variáveis relevantes para melhorar a precisão da predição nesse contexto clínico desafiador.

8 CONSIDERAÇÕES FINAIS

Com base nos resultados dos cenários apresentados, podemos concluir que a presença da SM está associada a um aumento significativo no RDC. As estatísticas descritivas mostraram que a população com SM apresentou uma média de RDC mais alta em comparação com a população geral.

No cenário 2, os modelos de regressão e classificação utilizando os marcadores de Framingham mostraram-se eficazes na predição da SM. O modelo de regressão Random Forest apresentou um bom ajuste aos dados, enquanto o modelo de classificação Random Forest demonstrou uma alta acurácia e métricas competitivas para ambas as classes.

No cenário 3, o modelo de classificação Random Forest otimizado apresentou uma alta acurácia na predição da SM. Suas métricas de precisão, recall e F1-score indicaram uma capacidade satisfatória de fazer previsões corretas e identificar corretamente os casos positivos.

Considerando os resultados dos cenários anteriores, o modelo de regressão e classificação Random Forest destacou-se como o mais eficaz e confiável na predição da SM e no RDC associado a ela. A combinação dos marcadores da SM com o marcador de sexo foi identificada como a abordagem mais simples e eficaz para a predição dessas condições de saúde.

No cenário 5, o modelo Random Forest mostrou uma boa capacidade de discriminação entre as classes positiva e negativa de RDC, com uma área significativa sob a curva ROC e métricas sólidas de precisão, recall e F1-score para ambas as classes.

No cenário 6, os resultados da matriz de confusão indicaram que as variáveis relacionadas à SM têm um impacto significativo na predição do RDC. No entanto, a taxa de acerto para a classe positiva foi baixa, indicando a necessidade de melhorias e ajustes nos modelos.

Portanto, com base nos resultados e avaliações dos cenários apresentados, o modelo de regressão e classificação Random Forest mostrou-se o mais adequado para a predição da SM e do RDC associado a ela. Sua utilização, juntamente com os marcadores da SM e outras variáveis relevantes, pode fornecer informações valiosas na identificação de indivíduos com maior risco e permitir intervenções e acompanhamento adequados e personalizados. No entanto, é importante realizar análises adicionais e validações em estudos futuros para garantir a robustez e generalização dos resultados obtidos.

8.1 LIMITAÇÕES ENCONTRADAS

Os três conjuntos de dados não possuíam todas as variáveis necessárias para o diagnóstico da SM e do FRS em conjunto. O conjunto de dados "MSWithInsulin.csv", disponível na plataforma data.world, foi extraído do National Center for Health Statistics (NHANES) e incluiu informações de pacientes diagnosticados com SM e pacientes sem a doença ao longo de um período de dez anos. Embora contivesse os marcadores diagnósticos da SM baseados no critério NCEP-ATP III, não incluía todos os marcadores usados no FRS, especificamente o marcador de Colesterol Total, o que comprometeu uma análise completa e detalhada.

Da mesma forma, o segundo conjunto de dados também não possuía todos os Marcadores Clínicos dos conjuntos de dados de Cleveland, Hungria, Suíça e VA Long Beach. O objetivo desse conjunto de dados era determinar se os pacientes tinham RDC. Apesar de ter apenas três marcadores para o diagnóstico da SM, ainda foi possível observar a importância da combinação de variáveis conforme utilizada, obtendo um resultado significativo quando combinado com idade e sexo juntamente com as outras três variáveis da SM. Com apenas cinco variáveis, os resultados foram estatisticamente significativos.

No que diz respeito ao terceiro conjunto de dados, o banco de dados de Framingham, a variável ausente era o volume da circunferência da cintura, que na realidade é a única variável que não contribui para o FRS, conforme afirmado pelo Framingham Heart Study (2021). Em vez disso, o Índice de Massa Corporal (IMC) foi utilizado como substituto.

8.2 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Para dar continuidade a esta pesquisa e desenvolver um escore que possa ser utilizado tanto em pacientes com e sem SM, é recomendável explorar o uso de técnicas de Aprendizado de Máquina (ML). Essa abordagem exigirá o envolvimento de uma equipe multidisciplinar composta por profissionais da área da saúde, como médicos, nutricionistas, educadores físicos e pacientes, além de especialistas em tecnologia, como analistas de dados e engenheiros de software. Será necessário estabelecer parcerias com hospitais e laboratórios clínicos, bem como reservar um tempo adequado para a realização das tarefas, incluindo a elaboração de consentimentos dos participantes. Além disso, é importante considerar o aspecto financeiro como um fator crítico para o desenvolvimento e implementação do escore proposto.

Como alternativa, caso os resultados sejam significativos, é recomendado o uso do critério diagnóstico da SM em conjunto com algumas variáveis do FRS para a população com

SM. Esses dados podem ser coletados por meio de entrevistas, anamnese e exames não invasivos e complexos. Seria ainda mais benéfico desenvolver um escore que utilize informações sem a necessidade de exames relativamente caros. Um escore de fácil utilização e acessível seria uma ferramenta importante para a prevenção de DCV, que são uma das principais causas de morte no mundo, de acordo com a Organização Mundial da Saúde.

REFERÊNCIAS

ABESO. Obesidade e síndrome metabólica, (Associação Brasileira para o Estudo da Obesidade e Síndrome Metabólica) Disponível: https://abeso.org.br/conceitos/obesidade-e-sindrome-metabolica/ acesso em 16 de abril de 2022. In: . [S.l.: s.n.], 2022.

ASSOCIATION, A. H. Metabolic syndrome, (What is Metabolic Syndrome) Disponível: https://www.heart.org/en/health-topics/metabolic-syndrome acesso em 22 de abril de 2023. In: . [S.l.: s.n.], 2020.

BITTENCOURT, G. **Inteligência Artificial. Ferramentas e Teorias**. 2. ed. Florianópolis: Editora da UFSC, 2001. 51 p. ISBN 85-328-0138-2.

BRANDÃO, e. a. I diretriz brasileira de diagnóstico e tratamento da síndrome metabólica. Arquivos Brasileiros de Cardiologia - Volume 84, Suplemento I, Abril 2005. In: . [S.l.: s.n.], 2005.

CARDIOLOGIA, S. B. de. Diretrizes brasileiras de diagnóstico e tratamento da síndrome metabólica. In: . [s.n.], 2019. Disponível em: <https://www.cardiol.br/wp-content/uploads/2019/10/Diretriz-Sindrome-Metabolica-2019.pdf>.

CODORNIZ, A. **DESCOBERTA DOS FATORES DE RISCO PARA DOENÇA CORONARIANA E MORTE SÚBITA**. 2000. Trabalho apresentado na Jornada da Sociedade Centro-Oeste de Cardiologia (SCOC) – A Cardiologia dos Primórdios à Porta do Terceiro Milênio (Descobertas, Revoluções, Mitos, Equívocos, Avanços).

DOMENICO, K. D. Causas e fatores de risco. **Cardiovascular News**, September 2021. Disponível em: <https://www.cardiovascularnews.com.br/causas-e-fatores-de-risco-da-sindrome-metabolica/>.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. d. **Inteligência Artificial: Uma abordagem de aprendizado de máquina**. [S.l.]: Grupo Gen-LTC, 2011.

FIGUEROA, A. Inteligência artificial na saúde. **SBIE**, v. 11, p. 32–43, 2020.

FRAMINGHAM Heart Study. **Framingham Heart Study**, 2022. Acessado em 30 de novembro de 2022. Disponível em: <https://www.framinghamheartstudy.org/>.

GIANNINI, S. **Aterosclerose e Dislipidemias**. 1ª. ed. São Paulo: BG Cultural Ed., 1998.

GOMES, P. C. T. **Introdução ao Aprendizado de Máquina**. 2019. <https://www.datageeks.com.br/aprendizado-de-maquina/>. Acesso em: 10 de novembro de 2022.

GRUNDY, S. M. Metabolic syndrome update. **Trends in Cardiovascular Medicine**, v. 26, n. 4, p. 364–373, 2016.

GRUS, J. **Data Science do Zero**. [S.l.]: Novatec Editora, 2019.

GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. [S.l.]: Novatec Editora, 2017.

HOYT, R. M. **Projeto de dados abertos com dados NHANES 2011-2012resumo - MSWithInsulin2 Dataset**. 2018. <https://data.world/rhoyt/librehealth-educational-ehr/workspace/project-summary?agentid=rhoyt&datasetid=librehealth-educational-ehr>.

HUANG, W.; YING, T. W.; CHIN, W. L. C.; BASKARAN, L.; MARCUS, O. E. H.; YEO, K. K.; KIONG, N. S. Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction. **Journal of Medical Systems**, Springer, v. 45, n. 7, p. 1–11, 2021.

II, G. F.; HURST, J. Atherosclerotic coronary heart disease: Historical perspective. In: **Hurst's The Heart**. 9th. ed. [S.l.]: McGraw-Hill, 1999. p. 1–30.

JANOSI, a. a. **Heart Disease**. 1988. UCI Machine Learning Repository. DOI: <10.24432/C52P4X>.

JUNQUEIRA, V.; BARRETO, S. M.; PASSOS, V. M. A.; ALMEIDA, M. d. C.; MARTINS, I. C.; RODRIGUES, A. P. S.; ARA'uJO, D. Prevalence of metabolic syndrome according to different criteria in a population-based study. **Arquivos brasileiros de endocrinologia metabologia**, Sociedade Brasileira de Endocrinologia e Metabologia, v. 55, n. 7, p. 195–202, 2011.

KALIL, C. C. et al. **Obesidade**. 1st. ed. São Paulo: Benvirá, 2017. 176 p. (Coleção Como cuidar). ISBN 978-85-5717-173-2.

KAPLAN, A. **Artificial Intelligence, Business and Civilization - Our Fate Made in Machines**. [S.l.]: Routledge, 2022. ISBN 9781032155319.

KITCHENHAM, a. a. **Guidelines for Performing Systematic Literature Reviews in Software Engineering**. [S.l.], 2007.

KLEINA. **A história da inteligência artificial**. 2018. Vídeo. Publicado em 23 de outubro de 2018 às 10:58. Disponível em: <https://www.tecmundo.com.br/mercado/135413-historia-inteligencia-artificial-video.htm>.

LIMA, J. **Computadores e Inteligência**. [S.l.]: Editora UFSC, 2017.

LOTUFO, P. A. Risco cardiovascular global: Novos conceitos sobre uma velha realidade. In: **Risco Cardiovascular Global**. São Paulo: Lemos Editorial, 1999. p. 31–43.

MAIGA, a. a. Comparison of machine learning models in prediction of cardiovascular disease using health record data, Universitas Atma Jaya Yogyakarta Yogyakarta, Indonesia 2021. acesso em 17 de abril de 2023. 2019.

METABOLOGIA, S. B. de Endocrinologia e. Tratamento da síndrome metabólica. **SBEM - Sociedade Brasileira de Endocrinologia e Metabologia**, 2022.

MILDENBERGER, T. Stephen marsland: Machine learning. an algorithmic perspective. **Statistical Papers**, Springer, v. 55, n. 2, p. 575, 2014.

Ministério da Saúde. Escore de risco global (erg) de framingham. In: . [s.n.], 13 de junho de 2023. Disponível em: <https://aps.saude.gov.br>.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicaç oes**, v. 1, n. 1, p. 32, 2003.

National Heart, Lung, and Blood Institute. **Framingham Heart Study**. 1948. <https://www.framinghamheartstudy.org/fhs-about/>. Acesso em: [22 de Janeiro de 2023].

NEGRÃO, C. E.; MATOS, L. D. N. J. d.; COSTA, A. R. d.; RAMALHO, A. C.; PIERIN, A. M. G.; SALLES, G. F.; GUIMARãES, J. I.; MONTEIRO, J. B. **I Diretriz Brasileira de Diagnóstico e Tratamento da Síndrome Metabólica**. 2005.

NOYES, J. **Artificial Intelligence with Common Lisp: Fundamentals of Symbolic and Numeric Processing**. Lexington, Massachusetts: D. C. Heath and Company, 1992. 9–25 p.

OBERMEYER, Z.; LEE, T. H. Lost in Thought–The Limits of the Human Mind and the Future of Medicine. **New England Journal of Medicine**, Massachusetts Medical Society, v. 377, n. 13, p. 1209–1211, 2017.

OMS. **https://www.who.int/publications/i/item/9789240037403**. 2021. Disponível em: https://www.who.int/publications/i/item/9789240031180. Acesso em: 08/11/2022.

OPAS, OMS. Oms publica primeiro relatório global sobre inteligência artificial na saúde e seis princípios orientadores para sua concepção e uso. In: . [s.n.], 2021. [Online; acessado em 08 de novembro de 2022]. Disponível em: <https://www.paho.org/pt/noticias/28-6-2021-oms-publica-primeiro-relatorio-global-sobre-inteligencia-artificial-na-saude-e>.

PEDREGOSA, a. a. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, 2011. Disponível em: <https://scikit-learn.org/stable/index.html>.

PINILLA, e. a. Desarrollo de un sistema para la predicción de ocurrencia del síndrome metabólico empleando técnicas de inteligencia artificial. **Universidad Autónoma de Bucaramanga UNAB - Facultad Ingeniería - Pregrado Ingeniería Biomédica**, 2021. Acesso em 11 de abril de 2022.

POLANCZYK, C. Prevenção de doenças cardiovasculares: avanços recentes e os próximos desafios. **Arquivos Brasileiros de Cardiologia**, v. 85, n. 1, p. 3–6, 2005.

PRECOMA. Atualização da diretriz de prevenção cardiovascular da sociedade brasileira de cardiologia – 2019) Arq Bras Cardiol. 2019; 113(4):787-891. In: . [S.l.: s.n.], 2019.

QUARESMA, J. **O que é Machine Learning? Introdução e Conceitos Básicos**. 2019. <https://medium.com/data-hackers/>. Acesso em: 11 nov. 2022.

SAÚDE, B. V. E. **Síndrome metabólica**. 2017. <https://www.bvsalud.org/temas-saude/?q=sindrome+metabolica>. Acesso em 01 de maio de 2023.

SOMITI. **Síndrome metabólica e doenças cardiovasculares**. 2020. <http://blog.somiti.org.br/sindrome-metabolica-e-doencas-cardiovasculares/>. Acessado em 30 de novembro de 2022.

SUDHA, P. N.; L, A.; S, R.; KASHYAP, S. A. A.; K, J. B. S. Machine learning techniques for cardiovascular risk score -prediction. In: **2021 IEEE Mysore Sub Section International Conference (MysuruCon)**. [S.l.: s.n.], 2021. p. 505–509.

TUNES, M. Terreno fértil para a inteligência artificial. **Revista Pesquisa FAPESP**, 2022. Disponível em: <https://brasilescola.uol.com.br/informatica/inteligencia-artificial.htm>.

WENG, S. F.; REPS, J.; KAI, J.; GARIBALDI, J. M.; QURESHI, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? **PloS one**, Public Library of Science, v. 12, n. 4, p. e0174944, 2017.

YANG, a. a. Comparison between metabolic syndrome and the framingham risk score as predictors of cardiovascular diseases among kazakhs in xinjiang. **BMC cardiovascular disorders**, Springer, v. 18, n. 1, p. 196, 2018.

APÊNDICE A – Algoritmo e Modelos: Código e Estruturas

A.1 EXPERIMENTOS NO CENÁRIO 1

Este trabalho faz parte de um estudo abrangente sobre o Risco Cardiovascular relacionado à Síndrome Metabólica. O objetivo principal é extratificar o risco da população utilizando os dados do Dataset MSWithInsulin.csv.

```
# Bibliotecas

import pandas as pd
import numpy as np
import seaborn as sns
from pandas._libs import index
from numpy.lib.shape_base import column_stack
from operator import add
from pandas.core.frame import DataFrame
from pandas.core import apply

# Importado o conjunto de dados
sm = pd.read_csv('MSWithInsulin.csv')
sm.head(1)

# Slecionado um subconjunto (das colunas) de um DataFrame:
# 6 Marcadores utilizados no Score de Framingham
df = sm[['age', 'sex', 'HDL', 'Hypertension', 'BloodGlucose',
'smoking', 'MetabolicSyndrome']]

df.head(2)

# Pre-processamento dos dados
# Remover valores ausentes
df = df.dropna()

# Funcaoo para calcular o Escore de Framingham com base nos
# dados de uma linha.
# Este calculo e baseado na tabela de Escore de Risco Global
(ERG) de Framingham do Ministerio da Saude,
# que foi fornecida atraves da imagem no seguinte
```

```
# link: https://linhasdecuidado.saude.gov.br/portal/obesidade
# -no-adulto/unidade-de-atencao-primaria/planejamento-
terapeutico/escore-risco-global-framingham/

# Atribuicao de pontos de acordo com o risco global,
# para homens

def framingham(linha):
  total_pontos = 0 # Comecamos com o total de pontos zerados.
  if linha['sex'] == 'Male':
# Vamos primeiro tratar da idade:
# Aqui estou acessando apenas o valor da coluna 'age' para a
# linha em questao.
    idade = linha['age']
# Dados de acordo com o Ministerio da Saude
    if   idade <= 34: total_pontos += 0
# Alterei os dados de acordo com o Ministerio da Saude
    elif idade <= 39: total_pontos += 2
    elif idade <= 44: total_pontos += 5
    elif idade <= 49: total_pontos += 6
    elif idade <= 54: total_pontos += 8
    elif idade <= 59: total_pontos += 10
    elif idade <= 64: total_pontos += 11
    elif idade <= 69: total_pontos += 12
    elif idade <= 74: total_pontos += 14
    elif idade >= 75: total_pontos += 15

#HDL-Colesterol
    hdlcol = linha['HDL']
    if hdlcol >= 60: total_pontos -= 2
    elif hdlcol <= 50: total_pontos -= 1
    elif hdlcol <= 45: total_pontos = 0
    elif hdlcol <= 35: total_pontos += 1
    elif hdlcol < 35: total_pontos += 2

#Pressao-arterial
    prearterial = linha['Hypertension']
    if prearterial == 0: total_pontos += 0
    elif prearterial == 1: total_pontos += 3
```

```
#Diabete, diagnostico seguindo a recomendacao da Abeso -
# https://abeso.org.br/conceitos/diabetes/ Informacao baseada
# no teste Jejum*
    diabete = linha['BloodGlucose']
    if diabete < 126: total_pontos += 0
    elif diabete >= 127: total_pontos += 3

#Tabagismo
tabagismo = linha['smoking']
if tabagismo == 'Never_smoker': total_pontos = 0
elif tabagismo == 'Current_every_day_smoker': total_pontos = 4

#Depois de tudo computado e so retornar o risco com
#base no total de pontos:
# Retornando aqui o valor da porcentagem
#de risco de ECV em 10 anos

    if total_pontos <= -3: return -1
    elif total_pontos == -2: return 1.1
    elif total_pontos == -1: return 1.4
    elif total_pontos == 0: return 1.6
    elif total_pontos == 1: return 1.9
    elif total_pontos == 2: return 2.3
    elif total_pontos == 3: return 2.8
    elif total_pontos == 4: return 3.3
    elif total_pontos == 5: return 3.9
    elif total_pontos == 6: return 4.7
    elif total_pontos == 7: return 5.6
    elif total_pontos == 8: return 6.7
    elif total_pontos == 9: return 7.9
    elif total_pontos == 10: return 9.4
    elif total_pontos == 11: return 11.2
    elif total_pontos == 12: return 13.2
    elif total_pontos == 13: return 15.6
    elif total_pontos == 14: return 18.4
    elif total_pontos == 15: return 21.6
    elif total_pontos == 16: return 25.3
    elif total_pontos == 17: return 29.4
```

```
    elif total_pontos >= 18: return 30

# Atribuicao de pontos de acordo com o
# risco global, para mulheres

  elif linha['sex'] == 'Female':
    total_pontos = 0 # Comeca com o total de pontos zerados.
# Vamos primeiro tratar da idade:
# Aqui estou acessando apenas o valor da coluna 'age'
# para a linha em questao.
    idade = linha['age']
# Alterei os dados de acordo com o Ministerio da Saude
    if   idade <= 34: total_pontos += 0
# Alterei os dados de acordo com o Ministerio da Saude
    elif idade <= 39: total_pontos += 2
    elif idade <= 44: total_pontos += 4
    elif idade <= 49: total_pontos += 5
    elif idade <= 54: total_pontos += 7
    elif idade <= 59: total_pontos += 8
    elif idade <= 64: total_pontos += 9
    elif idade <= 69: total_pontos += 10
    elif idade <= 74: total_pontos += 11
    elif idade >= 75: total_pontos += 12
    # Alterei os dados de acordo com o Ministerio da Saude

#HDL-Colesterol
    hdlcol = linha['HDL']
    if hdlcol >= 60: total_pontos -= 2
    elif hdlcol <= 50: total_pontos -= 1
    elif hdlcol <= 45: total_pontos = 0
    elif hdlcol <= 35: total_pontos += 1
    elif hdlcol < 35: total_pontos += 2

#Pressao-arterial
    prearterial = linha['Hypertension']
    if prearterial == 0: total_pontos += 0
    elif prearterial == 1: total_pontos += 3

# Diabete, diagnostico seguindo a recomendacao da
```

```
#Abeso - https://abeso.org.br/conceitos/diabetes/ Informacao
# baseada no teste Jejum*
    diabete = linha['BloodGlucose']
    if diabete < 126: total_pontos += 0
    elif diabete >= 127: total_pontos += 4

#Tabagismo
tabagismo = linha['smoking']
if tabagismo == 'Never_smoker': total_pontos = 0
elif tabagismo == 'Current_every_day_smoker': total_pontos = 3

# Depois de tudo computado e so retornar o risco com
# base no total de pontos:
    if total_pontos <= -2: return 0.5
    elif total_pontos == -1: return 1
    elif total_pontos == 0: return 1.2
    elif total_pontos == 1: return 1.5
    elif total_pontos == 2: return 1.7
    elif total_pontos == 3: return 2.0
    elif total_pontos == 4: return 2.4
    elif total_pontos == 5: return 2.5
    elif total_pontos == 6: return 3.3
    elif total_pontos == 7: return 3.9
    elif total_pontos == 8: return 4.5
    elif total_pontos == 9: return 5.3
    elif total_pontos == 10: return 6.3
    elif total_pontos == 11: return 7.3
    elif total_pontos == 12: return 8.6
    elif total_pontos == 13: return 10
    elif total_pontos == 14: return 11.7
    elif total_pontos == 15: return 13.7
    elif total_pontos == 16: return 15.9
    elif total_pontos == 17: return 18.5
    elif total_pontos == 18: return 21.6
    elif total_pontos >= 19: return 24.8
    elif total_pontos >= 20: return 28.5
    elif total_pontos >= 21: return 30.5

# Agora para criar a nova coluna, pode usar o metodo apply.
```

```
# Ele vai rodar linha a linha e executar a funcao acima:

# Classificacao de risco global, segundo Escore de Framingham.
# O nome da nova coluna e RiscoECV (Evento de cardiovascular)
# Categoria:
# Baixo = < 10%/ 10 anos
# - Moderado = 10 a 20%/ 10 anos e Alto = >20%/ 10 anos

df['RiscoECV'] = df.apply(lambda linha: framingham(linha), axis=1)

# Aplicar a funcao para criar a nova coluna 'Categoria'
# Categoria: Baixo = <10%/ 10 anos -
# Moderado = 10 a 20%/ 10 anos e Alto = >20%/ 10 anos
    def classificar_risco(risco):
        if risco < 10:
            return 'Baixo'
        elif risco >= 10 and risco <= 20:
            return 'Moderado'
        else:
            return 'Alto'

    df['Categoria'] = df['RiscoECV'].apply(classificar_risco)

# Funcao para mapear a categoria de risco para doencas
# correspondentes
    def mapear_doenca(categoria):
        if categoria == 'Baixo':
            return 'Sem␣doenca␣cardiovascular'
        elif categoria == 'Moderado':
            return 'Doenca␣cardiovascular␣intermediaria'
        else:
            return 'Doenca␣cardiovascular␣avancada'

# Aplicar a funcao para criar a nova coluna 'Doenca'
    df['Doenca'] = df['Categoria'].apply(mapear_doenca)

# Gerar uma tabela com a quantidade de pacientes de acordo
# com a categoria
```

```
import pandas as pd

tabela_categorias =
df['Categoria'].value_counts().reset_index()
tabela_categorias.columns = ['Categoria', 'Quantidade']
print(tabela_categorias)

# Criar Grafico de Barras
import matplotlib.pyplot as plt

# Ordenar as categorias em ordem crescente de
# acordo com o indice
contagem_categorias =
df['Categoria'].value_counts().sort_index()

# Mapear os valores da coluna "Categoria" para as descricoes
# correspondentes

categorias = ['Alto risco', 'Baixo risco', 'Moderado risco']

# Criar o grafico de barras com as cores correspondentes
plt.bar(categorias, contagem_categorias.values,
color=['#1f77b4', '#2ca02c', '#ff7f0e'])

# Definir o titulo do grafico e dos eixos
    plt.title('Contagem de Categorias')
    plt.xlabel('Categoria')
    plt.ylabel('Quantidade')

# Adicionar rotulos nas barras
    for i, v in enumerate(contagem_categorias.values):
        plt.text(i, v + 10, str(v), color='black',
        ha='center', fontweight='bold')

# Exibir o grafico
    plt.show()

# Calcular as estatisticas descritivas do "RiscoECV"
    estatisticas = df['RiscoECV'].describe()
```

```
    desc_stats = estatisticas.round(3)

# Exibir as estatisticas descritivas
    print(desc_stats)

# Contar a quantidade de uma determinada coluna
    df.sex.value_counts()

# Calculo para elaborar a estatistica de todas as
# tres populacoes
import matplotlib.pyplot as plt

# Calculo das estatisticas descritivas da populacao geral
    estatisticas_geral = df['RiscoECV'].describe()

# Calculo das estatisticas descritivas da populacao sem SM
    estatisticas_sem_sindrome = df1['RiscoECV'].describe()

# Calculo das estatisticas descritivas da populacao com SM
    estatisticas_com_sindrome = df2['RiscoECV'].describe()

# Calculo das diferencas percentuais
diff_media = (estatisticas_com_sindrome['mean'] -
estatisticas_sem_sindrome['mean']) /
estatisticas_sem_sindrome['mean'] * 100

# Exibicaoo das estatisticas descritivas
    print("Estatisticas da Populacao Geral:")
    print(estatisticas_geral)
    print()

    print("Estatisticas da Populacao sem Sindrome Metabolica:")
    print(estatisticas_sem_sindrome)
    print()

    print("Estatisticas da Populacao com Sindrome Metabolica:")
    print(estatisticas_com_sindrome)
    print()
```

```
print("Diferenca Percentual na Media do
RiscoECV entre as Populacoes:")
print(f"Diferenca: {diff_media:.2f}%")
print()

# Calculo das contagens de cada categoria na populacao sem SM
    count_sem_sindrome =
    df1['Categoria'].value_counts().reset_index()
    count_sem_sindrome.columns =
    ['Categoria', 'Quantidade']
    count_sem_sindrome['Porcentagem'] =
    count_sem_sindrome['Quantidade'] / len(df1) * 100

# Calculo das contagens de cada categoria na populacao com SM
    count_com_sindrome =
    df2['Categoria'].value_counts().reset_index()
    count_com_sindrome.columns =
    ['Categoria', 'Quantidade']
    count_com_sindrome['Porcentagem'] =
    count_com_sindrome['Quantidade'] / len(df2) * 100

# Distribuicao das Categorias nas Populacoes
import matplotlib.pyplot as plt

# Calculo das contagens de cada categoria na populacao sem SM
    count_sem_sindrome =
    df1['Categoria'].value_counts().reset_index()
    count_sem_sindrome.columns =
    ['Categoria', 'Quantidade']
    count_sem_sindrome['Porcentagem'] =
    count_sem_sindrome['Quantidade'] / len(df1) * 100

# Calculo das contagens de cada categoria na populacao com SM
    count_com_sindrome =
    df2['Categoria'].value_counts().reset_index()
    count_com_sindrome.columns =
    ['Categoria', 'Quantidade']
    count_com_sindrome['Porcentagem'] =
    count_com_sindrome['Quantidade'] / len(df2) * 100
```

```
# Plotagem do grafico de barras
    fig, ax = plt.subplots(figsize=(8, 6))

    ax.bar(count_sem_sindrome['Categoria'],
    count_sem_sindrome['Quantidade'], label='Sem SM')
    ax.bar(count_com_sindrome['Categoria'],
    count_com_sindrome['Quantidade'], label='Com SM')

    ax.set_xlabel('Categoria')
    ax.set_ylabel('Quantidade')
    ax.set_title('Distribuicao das Categorias nas Populacoes')
    ax.legend()
    plt.show()
```

O Cenário 1 representa uma etapa fundamental do estudo, onde são aplicadas técnicas específicas para analisar e quantificar o risco cardiovascular associado à Síndrome Metabólica. A estratificação do risco nos permite identificar diferentes grupos de indivíduos com base em seus perfis de risco.

A.2 CÓDIGO: EXPERIMENTOS NO CENÁRIO 2 - REGRESSÃO

Seleção do Melhor Modelo de Regressão

```
# Importar as bibliotecas necessarias
    import pandas as pd
    import numpy as np
    from sklearn.model_selection import train_test_split,
    GridSearchCV
    from sklearn.preprocessing import StandardScaler,
    OneHotEncoder
    from sklearn.impute import SimpleImputer
    from sklearn.compose import ColumnTransformer
    from sklearn.pipeline import Pipeline
    from sklearn.linear_model import LinearRegression
    from sklearn.linear_model import LogisticRegression
    from sklearn.tree import DecisionTreeRegressor
    from sklearn.ensemble import GradientBoostingRegressor
    from sklearn.ensemble import RandomForestRegressor
    from sklearn.svm import SVR
    from sklearn.neighbors import KNeighborsRegressor
    from sklearn.metrics import accuracy_score,
    mean_absolute_error,
    mean_squared_error, r2_score

# 1. Importando o DataSet
    sm = pd.read_csv('MSWithInsulin.csv')

# Pegar um subconjunto (das colunas) de um DataFrame:
# Marcadores do
# Escore de Framingham

d = sm[['age', 'sex', 'HDL', 'Hypertension', 'BloodGlucose',
'smoking', 'MetabolicSyndrome']]
d.head(2)

# Essa abordagem de codificacao e util quando se lida
# com variaveis categoricas que
```

```
# precisam ser representadas numericamente para uso
# em algoritmos de aprendizado de maquina.

# Bibliotecas
    from sklearn import preprocessing
    le = preprocessing.LabelEncoder()

    df = d.iloc[:,:]
    le.fit(df.sex)
    df['sex'] = le.transform(df.sex)
    le.fit(df.smoking)
    df['smoking'] = le.transform(df.smoking)

# Pre-processamento dos dados
# Remover valores ausentes
df = df.dropna()

# Separar as features e o target
X = df.drop(['MetabolicSyndrome'], axis=1)
y = df['MetabolicSyndrome']

# Criando pipelines para os modelos usados nesta pesquisa
pipelines = []
pipelines.append(('LR', Pipeline([('Regressor',
LogisticRegression())])))
pipelines.append(('DT', Pipeline([('Regressor',
DecisionTreeRegressor())])))
pipelines.append(('RF', Pipeline([('Regressor',
RandomForestRegressor())])))
pipelines.append(('GBM', Pipeline([('Regressor',
GradientBoostingRegressor())])))
pipelines.append(('SVM', Pipeline([('Regressor', SVR())])))
pipelines.append(('KNN', Pipeline([('Regressor',
KNeighborsRegressor())])))

# Dividir os dados em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

```
# Treinando e avaliaando os diferentes modelos cadastrados no pipelne
results = []
names = []
for name, pipeline in pipelines:
    pipeline.fit(X_train, y_train)
    y_pred = pipeline.predict(X_test)
    r2 = r2_score(y_test, y_pred)
    print(name, r2)
    results.append(r2)
    names.append(name)

# Exibir resultados em um grafico de barras
import matplotlib.pyplot as plt

fig, ax = plt.subplots()
ax.bar(names, results)
ax.set_xlabel('Modelos')
ax.set_ylabel('Coeficiente de determinacao (R2 ao quadrado)')
ax.set_ylim(bottom=0.0, top=1.0)
ax.set_title('Desempenho dos modelos')
plt.show()

# Escolher o melhor modelo
best_idx = np.argmax(results)
best_pipeline = pipelines[best_idx][1]
best_name = names[best_idx]

params = {}
if best_name == 'LR':
    params = {'Regressor__penalty': ['l1', 'l2',
            'elasticnet', 'none'],
              'Regressor__C': [0.01, 0.1, 1, 10, 100]}
elif best_name == 'DT':
    params = {'Regressor__criterion': ['poisson', 'friedman_mse',
                'absolute_error', 'squared_error'],
              'Regressor__max_depth': [3, 5, 7, 9]}
elif best_name == 'RF':
    params = {'Regressor__n_estimators': [100, 500, 1000],
              'Regressor__max_depth': [3, 5, 7, 9]}
```

```
elif best_name == 'GBM':
    params = {'Regressor__learning_rate': [0.01, 0.1, 1],
              'Regressor__n_estimators': [100, 500, 1000],
              'Regressor__max_depth': [3, 5, 7, 9]}
# Parametros para SVM
if best_name == 'SVM':
    params = {'Regressor__kernel': ['linear', 'poly',
              'rbf', 'sigmoid'],
              'Regressor__C': [0.01, 0.1, 1, 10, 100],
              'Regressor__gamma': ['scale', 'auto']}
# Parametros para KNN
elif best_name == 'KNN':
    params = {'Regressor__n_neighbors': [3, 5, 7, 9],
              'Regressor__weights': ['uniform', 'distance'],
              'Regressor__p': [1, 2]}

grid_search = GridSearchCV(best_pipeline, params,
cv=5, verbose=1)
grid_search.fit(X_train, y_train)

# Build a forest of trees from the training set (X, y).
model = RandomForestRegressor()
data = df
X = data.iloc[:,:-1]  #independent columns
y = data.iloc[:,-1]    #MetabolicSyndrome
model.fit(X,y)
print(model.feature_importances_) #use classe embutida
feature_importances de classificadores baseados em arvore
#plot grafico de importancias de recursos para melhor
visualizacao
feat_importances = pd.Series(model.feature_importances_,
index=X.columns)
feat_importances.nlargest(13).plot(kind='barh')
plt.show()

# Avaliar o melhor modelo otimizado
y_pred = grid_search.predict(X_test)
r2 = r2_score(y_test, y_pred)
print("Best␣Model:␣", best_name)
```

```
print("Best␣Model␣Parameters:␣", grid_search.best_params_)
print("Best␣Model␣Accuracy:␣", r2)
print("Best␣Model␣Regression:␣\n", r2_score(y_test, y_pred))

from sklearn.metrics import mean_absolute_error,
mean_squared_error, r2_score

# Avaliando o modelo para imprimir diferentes metricas:

def evaluate_model(y_true, y_pred):
    print("R2␣Score:␣", r2_score(y_true, y_pred))
    print("Mean␣Absolute␣Error:␣",
    mean_absolute_error(y_true, y_pred))
    print("Mean␣Squared␣Error:␣",
    mean_squared_error(y_true, y_pred))
    print("Root␣Mean␣Squared␣Error:␣",
    mean_squared_error(y_true, y_pred, squared=False))
# Avaliar o melhor modelo otimizado
y_pred = grid_search.predict(X_test)
evaluate_model(y_test, y_pred)

from tabulate import tabulate
import pandas as pd
from sklearn.metrics import mean_absolute_error,
mean_squared_error, r2_score

# Avaliar o melhor modelo otimizado
y_pred = grid_search.predict(X_test)

# Definir os resultados
results = pd.DataFrame({'Modelo': ['RF'],
    'R2␣Score': [r2_score(y_test, y_pred)],
    'Mean␣Absolute␣Error': [mean_absolute_error
     (y_test, y_pred)],
    'Mean␣Squared␣Error': [mean_squared_error
     (y_test, y_pred)],
    'Root␣Mean␣Squared␣Error': [mean_squared_error
     (y_test, y_pred,
     squared=False)]})
```

```
# Exibir a tabela formatada com Tabulate
print(tabulate(results, headers='keys', tablefmt='psql'))
```

A.3 CÓDIGO: EXPERIMENTOS NO CENÁRIO 2 - CLASSIFICAÇÃO

Seleção do Melhor Modelo de Classificação

```
# Importar as bibliotecas necessarias

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split,
    GridSearchCV
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier,
    GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score,
    classification_report

# 1. Importando o DataSet
    sm = pd.read_csv('MSWithInsulin.csv')
    sm.head(2)

# Pegar um subconjunto (das colunas) de um DataFrame:
# Marcadores usados para diagnostico da SM
df = sm[['age', 'sex', 'HDL', 'Hypertension',
         'BloodGlucose', 'smoking', 'MetabolicSyndrome']]

# Pre-processamento dos dados
# Remover valores ausentes
df = df.dropna()

# Essa abordagem de codificacao e util quando se lida
# com variaveis categoricas que
```

```
# precisam ser representadas numericamente para uso
# em algoritmos de aprendizado de maquina.
from sklearn import preprocessing
le = preprocessing.LabelEncoder()

df1 = df.iloc[:,:]
le.fit(df1.sex)
df1['sex'] = le.transform(df1.sex)
le.fit(df1.smoking)
df1['smoking'] = le.transform(df1.smoking)

# Separar as features e o target
X = df.drop(['MetabolicSyndrome'], axis=1)
y = df['MetabolicSyndrome']

# Criar pipelines para diferentes modelos
pipelines = []
pipelines.append(('LR', Pipeline([('classifier',
                    LogisticRegression())])))
pipelines.append(('DT', Pipeline([('classifier',
                    DecisionTreeClassifier())])))
pipelines.append(('RF', Pipeline([('classifier',
                    RandomForestClassifier())])))
pipelines.append(('GBM', Pipeline([('classifier',
                    GradientBoostingClassifier())])))
pipelines.append(('SVM', Pipeline([('classifier', SVC())])))
pipelines.append(('KNN', Pipeline([('classifier',
                    KNeighborsClassifier())])))

# Dividir os dados em treino e teste
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2, random_state=42)

# Treinar e avaliar diferentes modelos
results = []
names = []
for name, pipeline in pipelines:
    pipeline.fit(X_train, y_train)
    y_pred = pipeline.predict(X_test)
```

```
    acc = accuracy_score(y_test, y_pred)
    print(name, acc)
    results.append(acc)
    names.append(name)

# Escolher o melhor modelo
best_idx = np.argmax(results)
best_pipeline = pipelines[best_idx][1]
best_name = names[best_idx]

params = {}
if best_name == 'LR':
    params = {'classifier__penalty': ['l1', 'l2',
                'elasticnet', 'none'],
              'classifier__C': [0.01, 0.1, 1, 10, 100]}
elif best_name == 'DT':
    params = {'classifier__criterion': ['gini', 'entropy'],
              'classifier__max_depth': [3, 5, 7, 9]}
elif best_name == 'RF':
    params = {'classifier__n_estimators': [100, 500, 1000],
              'classifier__max_depth': [3, 5, 7, 9]}
elif best_name == 'GBM':
    params = {'classifier__learning_rate': [0.01, 0.1, 1],
              'classifier__n_estimators': [100, 500, 1000],
              'classifier__max_depth': [3, 5, 7, 9]}
elif best_name == 'SVM':
    params = {'classifier__C': [0.01, 0.1, 1, 10, 100],
              'classifier__kernel': ['linear', 'rbf',
              'poly', 'sigmoid']}
elif best_name == 'KNN':
    params = {'classifier__n_neighbors': [3, 5, 7, 9],
              'classifier__weights': ['uniform', 'distance'],
              'classifier__metric': ['euclidean', 'manhattan']}

grid_search = GridSearchCV(best_pipeline, params,
                           cv=5, verbose=1)
grid_search.fit(X_train, y_train)

import matplotlib.pyplot as plt
```

```
# Build a forest of trees from the training set (X, y).
model = RandomForestClassifier()
data = df
X = data.iloc[:,:-1]  #independent columns
y = data.iloc[:,-1]    #MetabolicSyndrome
model.fit(X,y)

# use classe embutida feature_importances de classificadores
# baseados em arvore

print(model.feature_importances_)

# plot grafico de importancias de recursos para
# melhor visualizacao
feat_importances = pd.Series(model.feature_importances_,
                                    index=X.columns)
feat_importances.nlargest(13).plot(kind='barh')
plt.show()

# Avaliar o melhor modelo otimizado
y_pred = grid_search.predict(X_test)
acc = accuracy_score(y_test, y_pred)
print("Best Model: ", best_name)
print("Best Model Parameters: ", grid_search.best_params_)
print("Best Model Accuracy: ", acc)
print("Best Model Classification Report: \n",
        classification_report(y_test, y_pred))

# Bibliotecas para plotar o grafico
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc

# Obter as probabilidades previstas para a classe positiva
y_prob = grid_search.predict_proba(X_test)[:, 1]

# Calcular a taxa de falsos positivos (FPR)
#e a taxa de verdadeiros positivos (TPR)

fpr, tpr, thresholds = roc_curve(y_test, y_prob)
```

```
# Calcular a area sob a curva ROC (AUC)
roc_auc = auc(fpr, tpr)

# Plotar a curva ROC
plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2,
         label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('Taxa de Falsos Positivos')
plt.ylabel('Taxa de Verdadeiros Positivos')
plt.title('Curva ROC')
plt.legend(loc="lower right")
plt.show()

# Biblioteca para plotar a Matriz
import seaborn as sns

# Gerar a matriz de confusao
cm = confusion_matrix(y_test, y_pred)

# Configurar o heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues")

# Definir rotulos e titulo
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")

# Exibir o grafico
plt.show()
```

A.4 CÓDIGO: EXPERIMENTOS NO CENÁRIO 5 - REGRESSÃO

Seleção do Melhor Modelo de Regressão com Todas as Variáveis

```
# Importar as bibliotecas necessarias
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import accuracy_score,
mean_absolute_error, mean_squared_error, r2_score

# carregar os dados em um dataframe
data = pd.read_csv("heart_statlog_cleveland_hungary_final.csv")
data.head(1)

X = data.drop('target', axis=1)
y = data['target']

# Criando pipelines para os modelos usados nesta pesquisa
pipelines = []
pipelines.append(('LR', Pipeline([('Regressor',
                    LogisticRegression())])))
pipelines.append(('DT', Pipeline([('Regressor',
                    DecisionTreeRegressor())])))
pipelines.append(('RF', Pipeline([('Regressor',
                    RandomForestRegressor())])))
pipelines.append(('GBM', Pipeline([('Regressor',
```

```
                    GradientBoostingRegressor())])))
pipelines.append(('SVM', Pipeline([('Regressor',
                    SVR())])))
pipelines.append(('KNN', Pipeline([('Regressor',
                    KNeighborsRegressor())])))

# Dividir conjunto de dados em treinamento e teste
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.3, random_state=42)

# Treinando e avaliaando os diferentes
# modelos cadastrados no pipelne
results = []
names = []
for name, pipeline in pipelines:
    pipeline.fit(X_train, y_train)
    y_pred = pipeline.predict(X_test)
    r2 = r2_score(y_test, y_pred)
    print(name, r2)
    results.append(r2)
    names.append(name)

# Exibir resultados em um grafico de barras
import matplotlib.pyplot as plt

fig, ax = plt.subplots()
ax.bar(names, results)
ax.set_xlabel('Modelos')
ax.set_ylabel('Coeficiente de determinacao (R2)')
ax.set_ylim(bottom=0.0, top=1.0)
ax.set_title('Desempenho dos modelos')
plt.show()

# Escolher o melhor modelo
best_idx = np.argmax(results)
best_pipeline = pipelines[best_idx][1]
best_name = names[best_idx]
```

```
params = {}
if best_name == 'LR':
    params = {'Regressor__penalty': ['l1', 'l2', 'elasticnet',
                                     'none'],
              'Regressor__C': [0.01, 0.1, 1, 10, 100]}
elif best_name == 'DT':
    params = {'Regressor__criterion': ['poisson', 'friedman_mse',
              'absolute_error', 'squared_error'],
              'Regressor__max_depth': [3, 5, 7, 9]}
elif best_name == 'RF':
    params = {'Regressor__n_estimators': [100, 500, 1000],
              'Regressor__max_depth': [3, 5, 7, 9]}
elif best_name == 'GBM':
    params = {'Regressor__learning_rate': [0.01, 0.1, 1],
              'Regressor__n_estimators': [100, 500, 1000],
              'Regressor__max_depth': [3, 5, 7, 9]}
# Parametros para SVM
if best_name == 'SVM':
    params = {'Regressor__kernel': ['linear', 'poly', 'rbf',
                                    'sigmoid'],
              'Regressor__C': [0.01, 0.1, 1, 10, 100],
              'Regressor__gamma': ['scale', 'auto']}
# Parametros para KNN
elif best_name == 'KNN':
    params = {'Regressor__n_neighbors': [3, 5, 7, 9],
              'Regressor__weights': ['uniform', 'distance'],
              'Regressor__p': [1, 2]}

grid_search = GridSearchCV(best_pipeline, params, cv=5,
                           verbose=1)
grid_search.fit(X_train, y_train)


# Build a forest of trees from the training set (X, y).
model = RandomForestRegressor()
df = data
X = data.iloc[:,:-1]  #independent columns
y = data.iloc[:,-1]    #MetabolicSyndrome
model.fit(X,y)
```

```
print(model.feature_importances_) #use classe
# embutida feature_importances de classificadores
# baseados em arvore
# plot grafico de importancias de recursos para
# melhor visualizacao
feat_importances = pd.Series(model.feature_importances_,
                             index=X.columns)
feat_importances.nlargest(13).plot(kind='barh')
plt.show()

from tabulate import tabulate
import pandas as pd
from sklearn.metrics import mean_absolute_error,
mean_squared_error, r2_score

# Avaliar o melhor modelo otimizado
y_pred = grid_search.best_estimator_.predict(X_test)
best_estimator = grid_search.best_estimator_

# Obter o nome do estimador
estimator_name = None
for name, estimator in best_estimator.named_steps.items():
    if hasattr(estimator, '__class__'):
        estimator_name = estimator.__class__.__name__
        break

# Definir os resultados
results = pd.DataFrame({'Modelo': [estimator_name],
                        'R2 Score': [r2_score(y_test, y_pred)],
                        'Mean Absolute Error':
                        [mean_absolute_error
                        (y_test, y_pred)],
                        'Mean Squared Error':
                        [mean_squared_error(y_test, y_pred)],
                         'Root Mean Squared Error':
                         [mean_squared_error(y_test,
                         y_pred, squared=False)]})

# Exibir a tabela formatada com Tabulate
```

```
print(tabulate(results, headers='keys', tablefmt='psql'))
```

Seleção do Melhor Modelo de Regressão com 5 Variáveis

Para este modelo, é necessário repetir os passos descritos no APÊNDICE A - Código: Experimentos no Cenário 5 - Regressão Seleção do Melhor Modelo de Regressão com 5 Variáveis, utilizando o seguinte procedimento para gerar o DataFrame com as variáveis desejadas. Esse procedimento é aplicável tanto para o caso em que são utilizadas 5 variáveis como para o caso em que são utilizadas 3 variáveis.

```
# Selecionar as variaveis da SM e FRS e formar um novo
# DataFrame

df = data[['age', 'sex', 'resting␣bp␣s', 'cholesterol',
           'fasting␣blood␣sugar', 'target']]
```

A.5 CÓDIGO: EXPERIMENTOS NO CENÁRIO 5 - CLASSIFICAÇÃO

Seleção do Melhor Modelo de Classificação com Todas as Variáveis

```
# Importar as bibliotecas necessarias
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier,
                             GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, classification_report

# 1. Importando o DataSet
```

```
data = pd.read_csv('heart_statlog_cleveland_hungary_final.csv')
data.head(2)

X = data.drop('target', axis=1)
y = data['target']

# Criar pipelines para diferentes modelos
pipelines = []
pipelines.append(('LR', Pipeline([('classifier',
                    LogisticRegression())])))
pipelines.append(('DT', Pipeline([('classifier',
                    DecisionTreeClassifier())])))
pipelines.append(('RF', Pipeline([('classifier',
                    RandomForestClassifier())])))
pipelines.append(('GBM', Pipeline([('classifier',
                    GradientBoostingClassifier())])))
pipelines.append(('SVM', Pipeline([('classifier',
                    SVC())])))
pipelines.append(('KNN', Pipeline([('classifier',
                    KNeighborsClassifier())])))

# Dividir os dados em treino e teste
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2, random_state=42)

#Treinar e avaliar diferentes modelos
results = []
names = []
for name, pipeline in pipelines:
    pipeline.fit(X_train, y_train)
    y_pred = pipeline.predict(X_test)
    acc = accuracy_score(y_test, y_pred)
    print(name, acc)
    results.append(acc)
    names.append(name)

# Exibir resultados em um grafico de barras
import matplotlib.pyplot as plt
```

```
fig, ax = plt.subplots()
ax.bar(names, results)
ax.set_xlabel('Modelos')
ax.set_ylabel('Coeficiente de determinacao (R2)')
ax.set_ylim(bottom=0.0, top=1.0)
ax.set_title('Desempenho dos modelos')
plt.show()

# Escolher o melhor modelo
best_idx = np.argmax(results)
best_pipeline = pipelines[best_idx][1]
best_name = names[best_idx]

params = {}
if best_name == 'LR':
    params = {'classifier__penalty': ['l1', 'l2', 'elasticnet',
                                      'none'],
              'classifier__C': [0.01, 0.1, 1, 10, 100]}
elif best_name == 'DT':
    params = {'classifier__criterion': ['gini', 'entropy'],
              'classifier__max_depth': [3, 5, 7, 9]}
elif best_name == 'RF':
    params = {'classifier__n_estimators': [100, 500, 1000],
              'classifier__max_depth': [3, 5, 7, 9]}
elif best_name == 'GBM':
    params = {'classifier__learning_rate': [0.01, 0.1, 1],
              'classifier__n_estimators': [100, 500, 1000],
              'classifier__max_depth': [3, 5, 7, 9]}
elif best_name == 'SVM':
    params = {'classifier__C': [0.01, 0.1, 1, 10, 100],
              'classifier__kernel': ['linear', 'rbf', 'poly',
              'sigmoid']}
elif best_name == 'KNN':
    params = {'classifier__n_neighbors': [3, 5, 7, 9],
              'classifier__weights': ['uniform', 'distance'],
              'classifier__metric': ['euclidean', 'manhattan']}

grid_search = GridSearchCV(best_pipeline, params, cv=5,
                           verbose=1)
```

```
grid_search.fit(X_train, y_train)


import matplotlib.pyplot as plt
# Build a forest of trees from the training set (X, y).
model = RandomForestClassifier()
df = data
X = df.iloc[:,:-1]  #independent columns
y = df.iloc[:,-1]    #MetabolicSyndrome
model.fit(X,y)
print(model.feature_importances_)
# use classe embutida feature_importances de
# classificadores baseados em arvore
# plot grafico de importancias de recursos
# para melhor visualizacao
feat_importances = pd.Series(model.feature_importances_,
                             index=X.columns)
feat_importances.nlargest(13).plot(kind='barh')
plt.show()

# Avaliar o melhor modelo otimizado
y_pred = grid_search.predict(X_test)
acc = accuracy_score(y_test, y_pred)
print("Best Model: ", best_name)
print("Best Model Parameters: ", grid_search.best_params_)
print("Best Model Accuracy: ", acc)
print("Best Model Classification Report: \n",
      classification_report(y_test, y_pred))

import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc

# Obter as probabilidades previstas para a classe positiva
y_prob = grid_search.predict_proba(X_test)[:, 1]

# Calcular a taxa de falsos positivos (FPR) e a
# taxa de verdadeiros positivos (TPR)
fpr, tpr, thresholds = roc_curve(y_test, y_prob)
```

```
# Calcular a area sob a curva ROC (AUC)
roc_auc = auc(fpr, tpr)

# Plotar a curva ROC
plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2,
label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('Taxa de Falsos Positivos')
plt.ylabel('Taxa de Verdadeiros Positivos')
plt.title('Curva ROC')
plt.legend(loc="lower right")
plt.show()

from sklearn.metrics import confusion_matrix
import seaborn as sns

# Gerar a matriz de confusao
cm = confusion_matrix(y_test, y_pred)

# Configurar o heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues")

# Definir rotulos e titulo
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")

# Exibir o grafico
plt.show()
```

Seleção do Melhor Modelo de Classificação com 5 Variáveis

Para este modelo, é necessário repetir os passos descritos no APÊNDICE A - Código: Experimentos no Cenário 5 - Seleção do Melhor Modelo de Classificação com 5 Variáveis, utilizando o seguinte procedimento para gerar o DataFrame com as variáveis desejadas. Esse procedimento é aplicável tanto para o caso em que são utilizadas 5 variáveis como para o caso

em que são utilizadas 3 variáveis.

```
# Selecionar as variaveis da SM e FRS e formar um novo
# DataFrame

df = data[['age', 'sex', 'resting␣bp␣s', 'cholesterol',
           'fasting␣blood␣sugar', 'target']]
```

A.6 CÓDIGO: EXPERIMENTOS NO CENÁRIO 6 - REGRESSÃO E CLASSIFICAÇÃO

Seleção do Melhor Modelo de Regressão com Todas as Variáveis

Para reproduzir os resultados obtidos neste experimento, é necessário importar as bibliotecas mencionadas abaixo e carregar o conjunto de dados específico utilizado nesta tarefa. Como o objetivo é recomendar um modelo com base nos passos adotados nos experimentos do cenário 6, siga as etapas a seguir:

1 - Importe as seguintes bibliotecas:

2 - Carregue o banco de dados específico para esta tarefa. Certifique-se de ter o arquivo de dados no formato apropriado (por exemplo, CSV) e defina o caminho correto para o arquivo. Por exemplo:

3 - Siga os passos adotados nos experimentos realizados no cenário 6, que incluem etapas como a separação dos atributos de entrada e do atributo alvo, pré-processamento dos dados, treinamento de modelos e avaliação da precisão. Lembre-se de adaptar o código de acordo com o seu ambiente e os detalhes específicos do conjunto de dados que você está utilizando.

```
# Importar as bibliotecas necessarias para os modelos de Regressao
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeRegressor
```

```
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import accuracy_score,
    mean_absolute_error, mean_squared_error, r2_score

# Carregar o conjunto de dados
data = pd.read_csv("framingham.csv")
data.head(1)
```

Para os modelos de classificação, os mesmos passos mencionados anteriormente devem ser adotados. Certifique-se de importar as seguintes bibliotecas necessárias para os modelos de classificação:

Além disso, você precisará carregar o conjunto de dados específico para esta tarefa. Certifique-se de ter o arquivo de dados no formato apropriado (por exemplo, CSV) e defina o caminho correto para o arquivo.

Após importar as bibliotecas e carregar o conjunto de dados, você pode seguir os mesmos passos adotados anteriormente para os modelos de regressão. Isso inclui etapas como a separação dos atributos de entrada e do atributo alvo, pré-processamento dos dados, treinamento dos modelos de classificação e avaliação da acurácia.

Certifique-se de ajustar o código de acordo com o seu ambiente e os detalhes específicos do conjunto de dados que você está utilizando.

```
# Importar as bibliotecas necessarias
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split,
    GridSearchCV
from sklearn.preprocessing import StandardScaler,
    OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier,
                             GradientBoostingClassifier
from sklearn.svm import SVC
```

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, classification_report

# Carregar o conjunto de dados
data = pd.read_csv("framingham.csv")
data.head(1)
```